RANDOMIZED EVALUATION WORKSHOP

The Abdul Latif Jameel Poverty Action Lab (J-PAL)
28-29 August 2014: Center for Environmental Economics and Policy and Africa Workshop, Pretoria

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

BOOKLET CONTENTS

Workshop Agenda

<table>
<tr><th></th><th>Thursday, 28 August</th><th>Friday, 29 August</th></tr>
<tr><td>9:00 - 9:30</td><td>Introduction to J-PAL & Impact Evaluation
Eunice Koranteng</td><td>Recap of Day 1</td></tr>
<tr><td>9:30-10:00</td><td rowspan="2">Lecture 1:
Why Evaluate?
Eunice Koranteng</td><td rowspan="2">Lecture 4:
Theory of Change
Eunice Koranteng</td></tr>
<tr><td>10:00 - 10:30</td></tr>
<tr><td>10:30 - 11:00</td><td colspan="2">Tea Break</td></tr>
<tr><td>11:00 - 11:30</td><td rowspan="2">Case Study:
Why Randomize?</td><td rowspan="2">Case Study:
Theory of Change</td></tr>
<tr><td>11:30 - 12:00</td></tr>
<tr><td>12:00 - 12:30</td><td rowspan="2">Lecture 2:
Why Randomize?
Emily Cupito</td><td rowspan="2">Lecture 5:
How to Randomize
Emily Cupito</td></tr>
<tr><td>12:30 - 13:00</td></tr>
<tr><td>13:00 - 13:30</td><td rowspan="2">Lunch</td><td rowspan="2">Lunch</td></tr>
<tr><td>13:30 - 14:00</td></tr>
<tr><td>14:00 - 14:30</td><td rowspan="2">Lecture 3:
Promising RCTs in Environmental Economics
Emily Cupito</td><td rowspan="2">Case Study:
How to Randomize</td></tr>
<tr><td>14:30 - 15:00</td></tr>
<tr><td>15:00 - 15:30</td><td></td><td></td></tr>
<tr><td>15:30 - 16:00</td><td rowspan="3">Group Work:
Choosing a Research Question</td><td rowspan="2">Group Work:
Research Design</td></tr>
<tr><td>16:00 - 16:30</td></tr>
<tr><td>16:30 - 17:00</td><td>Group Presentations & Closing Remarks</td></tr>
</table>

COURSE OVERVIEW

RANDOMIZED EVALUATION WORKSHOP

The Abdul Latif Jameel Poverty Action Lab (J-PAL) at the University of Cape Town, South Africa, and the Centre for Environmental Economics and Policy in Africa (CEEPA) 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's large academic research network to provide participants with practical guidance for understanding randomized evaluations, as well as share evidence from the body of randomized evaluations focusing on the environment.

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.

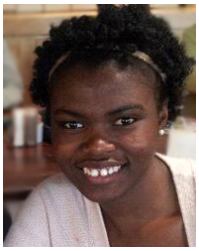
WORKSHOP METHODOLOGY

The workshop will incorporate the following:

- ***Lectures*** from experienced J-PAL 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.

WORKSHOP PRESENTERS

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.



EUNICE KORANTENG joined the J-PAL Africa office as a Research Associate in December 2012 and is currently working on projects in labor markets evaluating how a government intervention such as transportation subsidies for unemployed youth in South Africa effects job search and labor market movements. Eunice holds an Honors Degree in Finance and a Bachelor’s degree in Economics and Finance.

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'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

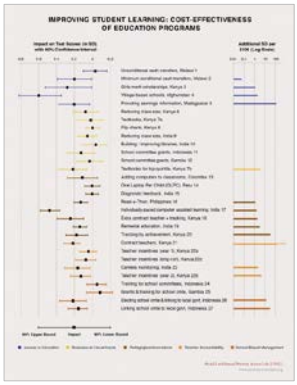
J-PAL's policy group produces policy publications to accompany the most successful or policy-relevant studies. Policy *briefcases* discuss a single study, while *bulletins* synthesize evidence from multiple studies and often accompany cost-electiveness analyses.

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.

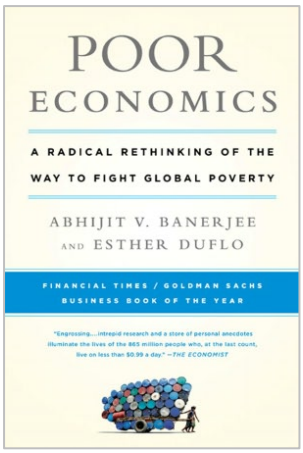


Full details of J-PAL's cost-effectiveness methodology, including assumptions on measuring costs and benefits, are included in the 2012 paper *Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries*, available at: *www.povertyactionlab.org/publication/cost-effectiveness*

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



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

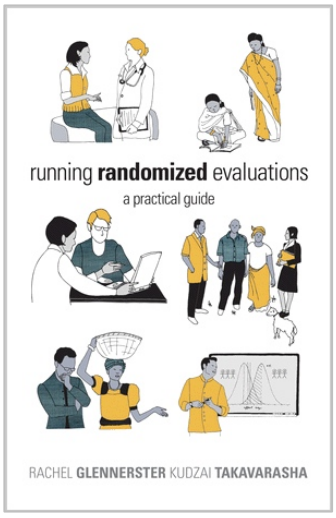
J-PAL's partner organizations include numerous research centers and program implementers. Key partner organizations are listed below. J-PAL's full partner database can be accessed from: *www.povertyactionlab.org/search/apachesolr_search?filters=type:partner*

- 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



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 EVALUATION

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

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.

- **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:

- **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.

- **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.

- **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.

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

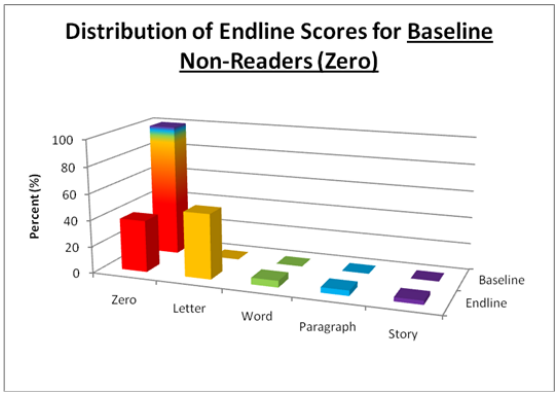
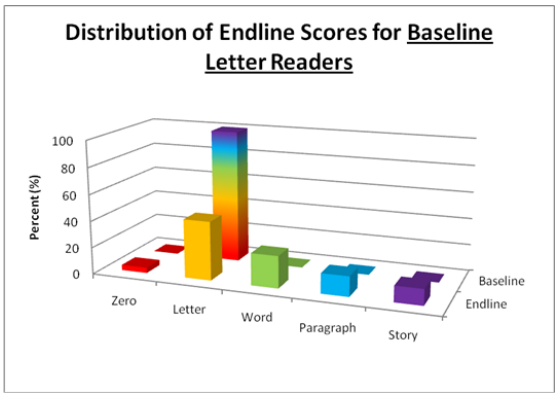


FIGURE 2



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.*

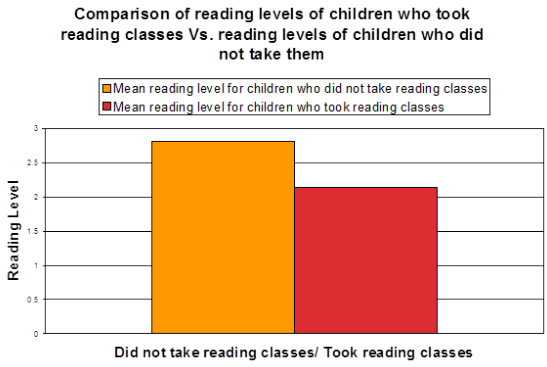
DISCUSSION TOPIC 1
Identifying evaluation

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

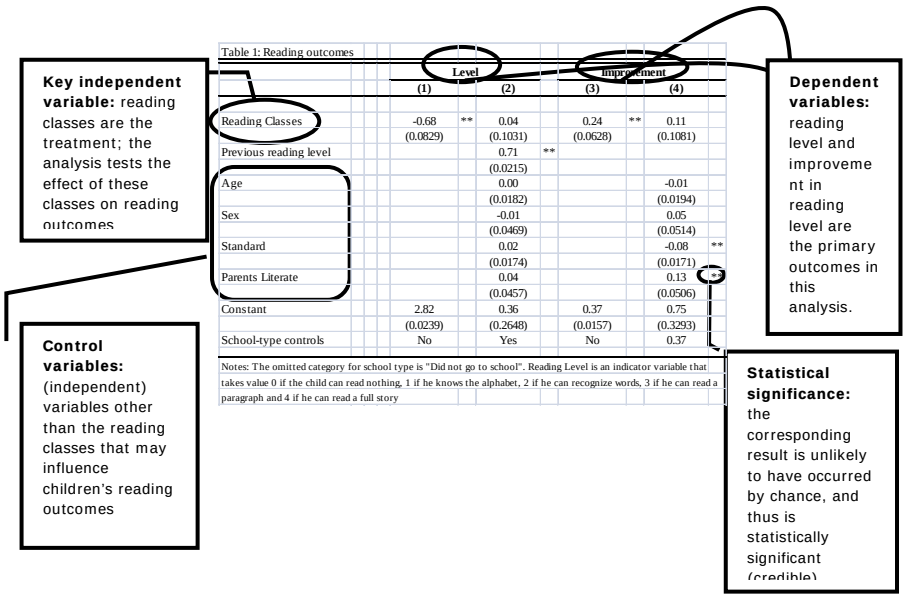
1. What type of evaluation does this news release imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

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.



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.

- **Measurable:** Research questions should have an outcome which can be measured. Researchers often must think about how to operationalize a concept which is not measured. For example, in order to measure if someone is depressed, researchers might ask specific questions about how often they cried in the last week as well as have them rate their sadness on a sliding scale.

In your groups, please discuss:

Discuss the following research questions. Analyze with your group whether they are a good or bad research question for an impact evaluation and identify ways in which you would make each question better.

- 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?
- 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 own research question here:

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Case Study 2: Women as Policymakers

Measuring the effects of political reservations

Thinking about measurement and outcomes

This case study is based on: Raghabendra Chattopadhyay and Esther Duflo, 2004a, "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India," *Econometrica* 72(5) 1409-1443.

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: 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 or failure.

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.

Introduction

India amended its federal constitution in 1992, devolving power over local development programs from the states to rural councils, or *gram panchayats* (village councils). The village councils now choose which development programs to undertake and how much of the budget to invest in them. The states are also required to reserve a third of village council seats and chairperson positions for women. In most states, the schedule on which different villages must reserve seats and positions is determined randomly. This creates the opportunity to rigorously assess the impact of quotas on politics and government: Do the policies differ when there are more women in government? Do the policies chosen by women in power reflect the policy priorities of women? Since randomization was part of the Indian government program itself, the evaluation planning centered on collecting the data needed to measure impact. The researchers then considered what data to collect and which data collection instruments to use.

Empowering the *panchayati raj*

Village councils, known locally as *panchayat*s, have a long tradition in rural India. Originally, *panchayats* were assemblies *(yat)* of five *(panch)* elders, chosen by the community, convened to mediate disputes between people or villages. In modern times village councils have been formalized into institutions of local self-government.

This formalization came about through the constitution. In 1992, India enacted the 73rd amendment, which directed the states to establish a three-tier *panchayati raj* system. The village council is the grassroot unit[1] of this system, with each council consisting of councilors elected every five years. The councilors elect from among themselves a chairperson called a *pradhan*. Decisions are made by a majority vote and the chairperson has no veto power. But as the only councilor with a full-time appointment, the chairperson wields effective power.

The 73rd amendment aimed to decentralize the delivery of public goods and services essential for development in rural areas. The states were directed to delegate the power to plan and implement local development programs to the village councils. Funds still come from the central government but are no longer earmarked for specific uses. Instead, the village council decides which programs to implement and how much to invest in them. As of 2005, Village Councils can chose programs from 29 specified areas, including welfare services (e.g., public assistance for widows, care for the elderly, maternity care, antenatal care, and child health) and public works (e.g., drinking water, roads, housing, community buildings, electricity, irrigation, and education).

1 Village councils, called *gram panchayats*, form the basic units of the *panchayat raj*. Village council chairs, elected by the members of the village council, serve as members of the block—subdistrict council (*panchayat samiti*). At the top of the system is the district council (*zilla parishad*) made up of the block–council chairs.

Empowering women in the *panchayati raj*

The village councils are large and diverse. In West Bengal, for example, each council represents up to 12 villages and up to 10,000 people who may vary by religion, ethnicity, caste, and, of course, gender.

Political voice varies by group identities drawn along these lines. If policy preferences vary by group identity and if the councilors' identities influence policy choices, then groups underrepresented in politics and government could be shut out as village councils could ignore those groups' policy priorities. There were fears that the newly empowered village councils would undermine the development priorities of traditionally marginalized groups such as women. To remedy this, the 73rd amendment included two mandates to ensure that investments reflected the needs of everyone in the village council.

The first mandate secures community input. If village council investments are to reflect a community's priorities, the councilors must first know what those priorities are. Accordingly, village councils are required to hold a general assembly every six months or every year to report on activities in the preceding period and to submit the proposed budget to the community for ratification. In addition, the chairpersons are required to set up regular office hours to allow constituents to formally request services and lodge complaints. Both requirements allow constituents to articulate their policy preferences.

The second mandate secures representation in the council for women. States are required to reserve at least a third of all council seats and chairperson positions for women. Furthermore, states must ensure that the seats reserved for women are "allotted by rotation to different constituencies in a *panchayat* (village council)" and that the chairperson positions reserved for women are "allotted by rotation to different *panchayat*s." In other words, they have to

ensure that reserved seats and chairperson positions rotate evenly within and across the village councils.

Randomized quotas in India: What can they teach us?

Your evaluation team has been entrusted with the responsibility to estimate the impact of quotas for women in the village councils. Your evaluation should address all dimensions in which quotas for women are changing local communities in India. What could these dimensions be? What data will you collect? What instruments will you use?

As a first step you want to understand all you can about the quota policy. What needs did it address? What are the pros and cons of the policy? What can we learn from it?

DISCUSSION TOPIC 1

Gender quotas in the village councils

1. What were the main goals of the village councils?
2. Women are underrepresented in politics and government. Only 10 percent of India's national assembly members are women, compared to 17 percent worldwide. Does it matter that women are underrepresented? Why and why not?
3. What were the framers of the 73rd amendment trying to achieve when they introduced quotas for women?

Gender quotas have usually been followed by dramatic increases in the political representation of women. Rwanda, for example, jumped from 24th place in the "women in parliament" rankings to first place (49 percent) after the introduction of quotas in 1996. Similar changes have been seen in Argentina, Burundi, Costa Rica, Iraq, Mozambique, and South Africa. Indeed, as of 2005, 17 of the top 20 countries in the rankings have quotas.

Imagine that your group is the national parliament of a country deciding whether to adopt quotas for women in the national parliament. Randomly divide your group into two parties, one against and one for quotas.

What data to collect

First, you need to be very clear about the likely impact of the program. It is on those dimensions that you believe will be affected that you will try to collect data. What are the main areas in which the quota policy should be evaluated? In which areas do you expect to see a difference as a result of quotas?

What are all the possible effects of quotas?

DISCUSSION TOPIC 2

Using a logical framework to delineate your intermediate and final outcomes of interest

1. Brainstorm the possible effects of quotas, positive, negative, and no effects.
2. What evidence would you collect to strengthen the case of those who are for or against quotas? For each potential effect on your list, also list the indicator(s) you would use for that effect. For example, if you say that quotas will affect political participation of women, the indicator could be "number of women attending the General Assembly."

Multiple outcomes are difficult to interpret, so define a hypothesis

Quotas for women could produce a large number of outcomes in different directions. For example, they may improve the supply of drinking water and worsen the supply of irrigation. Without an *ex ante* hypothesis on the direction in which these different variables should be affected by the quota policy, it will be very difficult to make sense of any result we find. Think of the following: if you take 500 villages and randomly assign them in your computer to a "treatment" group and a "control" group, and then run regressions to see whether the villages look

different along a hundred outcomes, would you expect to see some differences among them? Would it make sense to rationalize those results *ex post?*

The same applies to this case: if you present your report in front of the commission that mandated that you evaluate this policy, explaining that the quota for women changed some variables and did not change others, how should they interpret it? How will they know that these differences are not due to pure chance rather than the policy? It is necessary to present them with a clear hypothesis of how quotas are supposed to change policymaking, which will help you make predictions about which outcomes are affected.

DISCUSSION TOPIC 2, CONTINUED

3. What might be some examples of key hypotheses you could test? Pick one.

4. Which indicators or combinations of indicators would you use to test your key hypothesis?

Use a logical framework to delineate intermediate and final outcomes

A good way of figuring out the important outcomes is to lay out your theory of change; that is, to draw a logical framework linking the intervention, step by step, to the key final outcomes.

DISCUSSION TOPIC 2, CONTINUED

5. What are the steps or conditions that link quotas (the intervention) to the final outcomes?

6. Which indicators should you try to measure at each step in your logical framework?

7. Using the outcomes and conditions, draw a possible logical framework, linking the intervention and the final outcomes.

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Case Study 3: Extra Teacher Program

How to Randomize

This case study is based on the paper "Peer Effects and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," by Esther Duflo (MIT), Pascaline Dupas (UCLA), and Michael Kremer (Harvard)

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

Key vocabulary

Level of randomization: the level of observation (e.g., individual, household, school, village) at which treatment and comparison groups are randomly assigned.

Introduction

Confronted with overcrowded schools and a shortage of teachers, in 2005 the NGO International Child Support Africa (ICS) offered to help the school system of Western Kenya by introducing contract teachers in 120 primary schools. Under its two-year program, ICS provided funds to these schools to hire one extra teacher per school. In contrast to the civil servants hired by the Ministry of Education, contract teachers are hired locally by school committees. ICS expected this program to improve student learning by, among other things, decreasing class size and using teachers who are more directly accountable to the communities they serve. However, contract teachers tend to have less training and receive a lower monthly salary than their civil servant counterparts. Thus there was concern about whether these teachers were sufficiently motivated, given their compensation, or qualified, given their credentials.

What experimental designs could test the impact of this intervention on educational achievement? Which of these changes in the school landscape is primarily responsible for improved student performance?

Overcrowded schools

Like many other developing countries, Kenya has recently made rapid progress toward the Millennium Development Goal of universal primary education. Largely due to the elimination of school fees in 2003, primary school enrollment rose nearly 30 percent, from 5.9 million to 7.6 million between 2002 and 2005.

Without accompanying government funding, however, this progress has created its own set of new challenges in Kenya:

1. **Large class sizes:** Due to budget constraints, the rise in primary school enrollment has not been matched by proportional increases in the number of teachers. (Teacher salaries already account for the largest component of educational spending.) The result has been very large class sizes, particularly in lower grades. In a sample of schools in Western Kenya, for example, the average first grade class in 2005 had 83 students. This is concerning because it is believed that small classes are most important for the youngest students, who are still acclimating to the school environment. The Kenyan National Union of Teachers estimates that the country needs an additional 60,000 primary school teachers in addition to the existing 175,000 in order to reach all primary students and decrease class sizes.

2. **Teacher absenteeism:** Further exacerbating the problem of high pupil-teacher ratios, teacher absenteeism remains high, reaching nearly 20 percent in some areas of Kenya.

   There are typically no substitutes for absent teachers, so students simply mill around, go home, or join another class, often in a different grade. Small schools, which are prevalent in rural areas of developing countries, may be closed entirely as a result of teacher absence. Families have to consider whether school will even be open when deciding whether or not to send their children to school. An obvious result is low student attendance—even on days when the school is open.

3. **Heterogeneous classes:** Classes in Kenya are also very heterogeneous, with students varying widely in terms of school preparedness and support from home.

   Grouping students into classes sorted by ability (known as tracking, or streaming) is controversial among academics and policymakers. On one hand, if teachers find it easier to teach a homogeneous group of students, tracking could improve school effectiveness and test scores. Many argue, on the other hand, that if students learn in part from their peers, tracking could disadvantage low-achieving students while benefiting high-achieving students, thereby exacerbating inequality.

4. **Scarce school materials:** Because of the high costs of educational inputs and the rising number of students, educational resources other than the teacher are stretched, and in some cases up to four students must share one textbook. Additionally, an already overburdened infrastructure deteriorates faster when forced to serve more children.

5. **Low completion rates:** As a result of these factors, completion rates are very low in Kenya, with only 45.1 percent of boys and 43.3 percent of girls completing the first grade.

All in all, these issues pose a new challenge to the community: how to ensure minimum quality of education given Kenya's budget constraints.

What are contract teachers?

Governments in several developing countries have responded to similar challenges by staffing unfilled teaching positions with locally hired contract teachers who are not civil service employees. There are four

main characteristics of contract teachers: they are (1) appointed on annual renewable contracts, with no guarantee of renewed employment (unlike regular civil service teachers); (2) often less qualified than regular teachers and much less likely to have a formal teacher training certificate or degree; (3) paid lower salaries than those of regular teachers (typically less than a fifth of the salaries paid to regular teachers); and (4) more likely to be from the local area where the school is located.

Are contract teachers effective?

The increasing use of contract teachers has been one of the most significant policy innovations in providing primary education in developing countries, but it has also been highly controversial. Supporters say that using contract teachers is an efficient way of expanding education access and quality to a large number of first-generation learners. Knowing that the school committee's decision of whether or not to rehire them the following year may hinge on performance, contract teachers are motivated to try harder than their tenured government counterparts. Contract teachers are also often more similar to their students geographically, culturally, and socioeconomically.

Opponents argue that using underqualified and untrained teachers may staff classrooms, but will not produce learning outcomes. Furthermore, the use of contract teachers de-professionalizes teaching, reduces the prestige of the entire profession, and reduces motivation of all teachers. Even if it helps in the short term, it may hurt efforts to recruit highly qualified teachers in the future.

While the use of contract teachers has generated much controversy, there is very little rigorous evidence regarding the effectiveness of contract teachers in improving student learning outcomes.

The Extra Teacher Program randomized evaluation

In January 2005, ICS Africa initiated a two-year program to examine the effect of contract teachers on education in Kenya. Under the program, ICS gave funds to 120 local school committees to hire one extra contract teacher to teach an additional first grade class. The purpose of this intervention was to address three challenges: class size, teacher accountability, and heterogeneity of ability. The evaluation was designed to measure the impact of class-size reductions, the relative effectiveness of contract teachers, and how tracking by ability would impact both low- and high-achieving students.

Addressing multiple research questions through experimental design

Different randomization strategies may be used to answer different questions. What strategies could be used to evaluate the following questions? How would you design the study? Who would be in the treatment and control groups, and how would they be randomly assigned to these groups?

DISCUSSION TOPIC 1

Testing the effectiveness of contract teachers

1. What is the relative effectiveness of contract teachers versus regular government teachers?

DISCUSSION TOPIC 2

Looking at more general approaches to improving education

1. What is the effect of grouping students by ability on student performance?
2. What is the effect of smaller class sizes on student performance?

DISCUSSION TOPIC 3

Addressing all questions with a single evaluation

1. Could a single evaluation explore all of these issues at once?
2. What randomization strategy could do so?

GROUP WORK 2:

RESEARCH DESIGN

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

1. How will your intervention be administered? What will participants be required to do to participate?

2. What is the target population for which your project will focus? Be specific!

3. 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)?

4. What is your unit of randomization?

5. What outcomes are you going to analyze?

6. What type of data will you collect? From whom? On what timeline?

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Group Work: 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

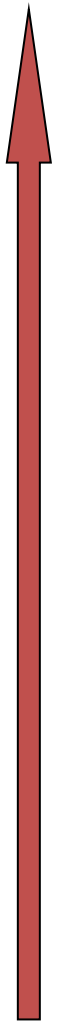
Discuss the theory of change of this intervention. In the table on the next page, fill in the inputs, outputs, outcomes, and impact. (Check the Key Vocabulary section to see definitions for theory of change, outputs, and outcomes.)

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				