ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

European Bank
for Reconstruction and Development

Evaluating Social Programs

June 22 – 26, 2009

J-PAL Executive Education at the European Bank for Reconstruction and Development.

Massachusetts Institute of Technology

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

European Bank
for Reconstruction and Development

J-PAL/EBRD Executive Education Course

June 22 – June 26, 2009
London, UK
EBRD, Randa Smine room 1206/1208

DAY 1 (June 22, 2009)

9:15 – 9:30	Welcome
9:30 – 11:00	Lecture 1: What is an Evaluation?
	Speaker: Rachel Glennerster (MIT, J-PAL)
11:00 – 12:00	Group Work and Case Study 1
12:00 – 13:30	Lunch
13:30 – 15:00	Lecture 2: Why Randomize?
	Speaker: William Parienté (Paris School of Economics, J-PAL)
15:00 – 16:30	Group Work
16:30 – 17:30	Case Study 2

DAY 2 (June 23, 2009)

9:00 – 10:30	Group Work / Case Study 2
10:30 – 12:00	Lecture 3: How to Randomize (Part 1)?
	Speaker: Greg Fischer (LSE)
12:00 – 13:30	Lunch
13:30 – 15:00	Lecture 4: How to Randomize (Part 2)?
	Speaker: Dean Karlan (Yale)
15:00 – 16:30	Group Work
16:30 – 18:00	Case Study 3

DAY 3 (June 24, 2009)

8:00 – 10:30	Group Work / Case Study 3
10:30– 12:00	Lecture 5: Measurements and Outcomes
	Speaker: Oriana Bandiera (LSE)
12:00 – 13:30	Lunch
13:30 – 15:00	Lecture 6: Power Calculations
	Speaker: Bruno Crépon (CREST)
15:00 – 16:30	Group Work
16:30 – 18:00	Case Study 4

Massachusetts Institute of Technology

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

DAY 4 (June 25, 2009)

8:00 – 10:30	Group Work / Case Study 4
10:30 – 00:00	Lecture 7: Threats to Validity
	Speaker: *Luc Behaghel (Paris School of Economics)*
12:00 – 13:00	Lunch
13:00 – 14:30	Lecture 8: Data Analysis
	Speaker: *Imran Rasul (UCL)*
14:30 – 18:00	Group Work / Preparation for Presentations

DAY 5 (June 26, 2009)

9:00 – 12:15	Group Presentations
12:15 – 13:15	Lunch
13:15 – 15:30	Group Presentations
15:30 – 16:00	Concluding Remarks

Lecturer Bios

Oriana Bandiera is an Associate Professor of Economics at the London School of Economics. She specializes in the design of field experiments to evaluate how individual behavior is shaped by monetary incentives and social relationships. Recent work covers field experiments on incentives for pro-social tasks in Zambia and the randomized evaluation of large scale poverty reduction and female empowerment interventions in Bangladesh, Uganda and Tanzania.

Luc Behaghel is a researcher at the Laboratoire d'economie appliquee (LEA-INRA) at the Paris School of Economics. His research interests include the impact of an aging population on labor market outcomes, technical change, and the evaluation of labor market and rural development policies. He is working on several, large-scale randomized evaluations of programs aimed at counseling the unemployed in France, mentoring high school students to help them choose their career path, as well as, a program designed to help integrate the parents of middles school students in their child's education.

Bruno Crépon is a researcher at CREST and a professor at ENSAE and Ecole Polytechnique in Paris, France. A focus of his research is policy evaluation with special attention to labor market policies, training programs and micro-credit in developed countries. He his currently conducting randomized evaluations in France of counselling schemes focused on the unemployed and welfare recipients. His is also conducting field experiments in Morocco on microcredit and entrepreneurship.

Greg Fischer is a Lecturer (Assistant Professor) of Economics at the London School of Economics. His research focuses on corporate finance, entrepreneurship, and financial innovation in developing countries. Prior to returning to academia, Greg worked for nine years in the private equity and venture capital arms of Morgan Stanley and Centre Partners, an affiliate of Lazard. His current work includes the randomized evaluations of a business training program in the Dominican Republic and the pricing of new water treatment technology in Ghana.

Rachel Glennerster joined the Abdul Latif Jameel Poverty Action Lab at MIT as Executive Director in 2004. She acted as Technical Assistant to the UK Executive Director of the IMF and World Bank focusing on loans to Russia and the former Soviet Union before joining the IMF staff in 1997. At the IMF she assisted countries affected by the Kosovo crisis, helped negotiate a major debt relief package for Mozambique, and helped design and implement reforms to the International Financial System in the aftermath of the Asian Financial Crisis. Her current research includes evaluations of public health and education interventions in India, community-driven development in Sierra Leone, and ways to empower adolescent girls in Bangladesh.

Dean Karlan is President and Founder of Innovations for Poverty Action, an Assistant Professor of Economics at Yale University, an Affiliate at the Abdul Latif Jameel Poverty Action Lab at the Massachusetts Institute of Technology, and an Affiliate of the Bureau for Research and Economic Analysis of Development (BREAD). His research focuses on microeconomic issues of public policies and poverty. He studies the effectiveness of particular policies to fight poverty and the relevance of economic theories of individual decision-making. Much of his work uses behavioral economic insights and approaches to examine economic and policy issues relevant in developing countries as well as in domestic charitable fundraising and political participation.

William Parienté is a post-Doctoral fellow at the Paris School of Economics. He wrote his dissertation on the analysis of credit demand and the evaluation of policies improving access to credit in three countries: Serbia, Brazil and Morocco, where he worked before joining J-PAL in 2006. His current research focuses on access to credit, entrepreneurship, poverty, and health issues. He is currently working on several randomized evaluations in Morocco, Pakistan and France.

Imran Rasul is a Reader (Associate Professor) in the Department of Economics at University College London. His areas of interest focus on labor economics and household economics. Previous and ongoing projects include studying whether and how a household's behavior is influenced by the presence and characteristics of its extended family and the evaluation of a female adolescent empowerment program in rural Uganda and Tanzania.

2009 J-PAL Executive Education at EBRD

First Name	Last Name	Country	Title	Organization	Email
Tsholofelo	Adelekan	South Africa	Deputy Director Monitoring and Evaluation	Department of Social Development	Tsholoa@socdev.gov.za
Lenka	Benova	Slovakia	Researcher	Social Research Center, The American University in Cairo	lbenova@aucegypt.edu
Claire	Bernard	Morocco	Reaserch Assistant	J-PAL	clairebernard@hotmail.de
Amy	Challen	United Kingdo	Research Officer	Centre for Economic Performance, London School of Economics	a.r.challen@lse.ac.uk
Amirah	El-Haddad	Egypt	Assistant Professor	Faculty of Economics and Political Sciences, Cairo University	amirah.elhaddad@gmail.com
Raissa	Fabregas Robles	Chile	Reaserch Assistant	J-PAL	raissa.fabregas-robles-gil@st-annes.ox.ac.uk
Hans Rudolf	Felber	Switzerland	Scientific collaborator	NADEL	felber@nadel.ethz.ch
Sefira	Fialkoff	Kenya	Reaserch Assistant	IPA	sfialkof@ucsc.edu
Mark	Fiorello	Sierra Leone	Reaserch Assistant	J-PAL	mark.fiorello@gmail.com
Véronique	Fradin	France	Reaserch Assistant	J-PAL	veronique.fradin@sciences-po.org
Hana	Freymiller	Ghana	Reaserch Assistant	IPA	HanaSF@alum.wellesley.edu
Rob	Fuller	Ghana	Reaserch Assistant	IPA	rfuller@poverty-action.org
Ugo	Gentilini	Italy	Policy adviser	United Nations World Food Programme (WFP)	Ugo.Gentilini@wfp.org
Maria	Gheorghiu	Romania	Executive Director	Asociatia Ovidiu Romm	maria@ovid.ro
Hélène	Giacobino	France	Director of Strategy and Development, J-PAL Europe	J-PAL	hgiacobino@povertyactionlab.org
Tharanga	Godallage	Sri Lanka	Monitoring Evaluations and Learning Coordinator	OXFAM GB	tharangahgo@hotmail.com
Danielle	Greco	France	Head of Monitoring and Evaluation	Pole emploi	danielle.greco@pole-emploi.fr
Sherri	Hansen	USA	Research Assistant	The William and Flora Hewlett Foundation	shansen@hewlett.org
Sarah	Kabay	Uganda	Reaserch Assistant	IPA	skabay@poverty-action.org
Daniel	Kariuki	U.S.	Regional M&E manager	TechnoServe	dkariuki@tns.org
Ryan	Knight	Ghana	Reaserch Assistant	IPA	ryangknight@gmail.com
Tomoki	Kobayashi	Japan	Assistant Director, Evaluation Division 1, Evaluat	Japan International Cooperation Agency (JICA)	kobayashi.tomoki@jica.go.jp
Laure	Le Douarec	France	Global Diversity Manager	Schneider Electric	laure@2d4b.com
Clémence	Le Roy	France	Monitoring and Evaluation Officer	Pole emploi	clemence.le-roy@pole-emploi.fr
Dimitris	Mavridis	India	Reaserch Assistant	J-PAL	mavridis.dimitris@gmail.com
Richard	McDowell	USA	Reaserch Assistant	J-PAL	richardmcdowell@gmail.com
Lilit	Melikyan	Republic of Ar	Monitorign and Evaluation Manager	Water and Sanitation for the Urban Poor	lmelikyan@wsup.com
Bastien	Michel	Kenya	Reaserch Assistant	IPA	bmichel@rip.ens-cachan.fr
Alex	Nisichenko	Ghana	Reaserch Assistant	IPA	anisichenko@poverty-action.org
Moleboheng	Ntene	South Africa	Economist	Department of Trade and Industry	MNtene@thedti.gov.za
Catherine	Palpant	France	Advisor for European and international issues	Cabinet of the High Commissioner for active inclusion against poverty	cathernine.palpant@pm.gouv.fr
Victor	Pouliquen	Morocco	Reaserch Assistant	J-PAL	vpouliquen@povertyactionlab.org
Alexia	Pretari	Morocco	Reaserch Assistant	J-PAL	alexia.pretari@gmail.com
Justyna	Pytkowska	Poland	Research Specialist	Fundacja Centrum Organizacji Pozyczkowych (Microfinance Centre)	justyna@mfc.org.pl
Amanda	Satterly	Tanzania	Tanzania M&E manager	TechnoServe	amanda.satterly@tnstanzania.org
Richard	Sawyer	Mongolia	Reaserch Assistant	IPA	rsawyer4@jhu.edu
Chhavi	Sharma	India	Project Manager	Freeplay Foundation	csharma@freeplayfoundation.org
Kentaro	Toyama	USA	Assistant Managing Director	Microsoft Research India	kentoy@microsoft.com
Alexandra	Zoueva	USA/Russia	Portfolio Analyst	The Childrens Investment Fund Foundation (CIFF)	Sasha@ciff.org
Group Leaders					
Francesco	Avvisati	France	Ph.D. Candidate	J-PAL, Paris School of Economics	nina.guyon@gmail.com
Nik	Buehren	UK	Ph.D. Candidate	UCL	n.buehren@ucl.ac.uk
Selim	Gulisci	UK	Ph.D. Candidate	LES	S.Gulesci@lse.ac.uk
Nina	Guyon	France	Ph.D. Candidate	J-PAL, Paris School of Economics	francesco.avvisati@gmail.com
Elizabeth	Linos	USA	Ph.D. Candidate	J-PAL, Harvard	elinos@povertyactionlab.org
Sanchari	Roy	UK	Ph.D. Candidate	LES	S.Roy2@lse.ac.uk
Juliette	Seban	France	Ph.D. Candidate	J-PAL, Paris School of Economics	jseban@povertyactionlab.org
Dan	Stein	UK	Ph.D. Candidate	LES	D.Stein@lse.ac.uk
Lecturers					
Oriana	Bandiera	UK	Reader (Associate Professor)	LES	O.Bandiera@lse.ac.uk
Luc	Behaghel	France	Associate Professor	Paris School of Economics	luc.behaghel@ens.fr
Bruno	Crepon	France	Professor	CREST, ENSAE, Ecole Polytechnique	bruno.crepon@ensae.fr
Greg	Fischer	UK	Lecturer (Assistant Professor)	LES	G.Fischer@lse.ac.uk
Rachel	Glennerster	USA	Executive Director	J-PAL, MIT	rglenner@mit.edu
Dean	Karlan	USA	Assistant Professor of Economics	Yale, IPA, J-PAL	dean.karlan@yale.edu
William	Pariente	France	Post Doctoral Fellow	Paris School of Economics	william.pariente@parisschoolofeconomics.eu
Imran	Rasul	UK	Reader (Associate Professor)	UCL	i.rasul@ucl.ac.uk

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

European Bank
for Reconstruction and Development

Group Work Instructions

Groups are assigned to the follow locations

Group 1, Room 808
Group 2, Room 809
Group 3, Room 811
Group 4, Room 812
Group 5, Room 814
Group 6, Room 815
Group 7, Randa Smine Room 1206/1208
Group 8, Randa Smine Room 1206/1208

You will be assigned to groups of 5-6 people. We will do our best to ensure that each group includes participants with a range of different experiences but some common areas of interest. You will carry out two types of activities within these groups:

i) Casework and discussions

ii) Preparation of group proposal

Casework and Discussions

Each case covers a specific set of topics which are the subject for the lectures for each day of the course. The cases provide background on one (or in some cases two) specific evaluations which will be referred to in the lectures. In addition, each case includes discussion topics designed to get you thinking about the issues prior to the lectures. Some of the cases also include exercises for you to complete. You will be provided with Excel files containing these exercises at the start of the “group work” sessions. You will be expected to read the relevant case, go through the discussion topics, and complete the exercises before the related lecture on the case.

It is very important that you come to the case discussion having read the case as there is no time to read the case and work through the questions in the time allocated.

Group Proposal

Each group will—over the course of the week—work on a proposal for an evaluation on a topic of their choice. Different aspects of evaluation will be covered in the lectures and the casework, and these should be reflected in the group proposal. On Saturday, each group will present their proposal and receive comments from the other participants and the lecturers. This is an ideal time to get feedback on an evaluation you may be planning.

The output for the project will be a 20-minute presentation (with an additional 10 minutes for questions and feedback).

The presentation should cover the following issues:

i) The objective and rationale of the evaluation—what is the question you are asking and why is it important or interesting?

ii) Randomization design—how will the treatment and control groups be determined, and at what level will the randomization take place?

iii) Measurement issues—how will you measure whether the program is a success? On what variables will data be collected? How will it be collected? In addition to final outcome measures, will you be collecting data on the mechanism by which the program works? If so, what data will you collect on this?

iv) What magnitude of effect will you be trying to detect? What is the sample size you will be using? Why is this the correct sample size?

v) What are the risks to the integrity of the evaluation? How will you seek to minimize these?

vi) How will the data be analyzed?

vii) To what use will you put the results? How will the results impact future policy/programs?

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Massachusetts Institute of Technology

Courtesy of Flickr user theocean

Case 1: Get out the vote

Do phone calls to encourage voting work? Why randomize?

This case study is based on "Comparing Experimental and Matching Methods Using a Large-Scale Field Experiment on Voter Mobilization," by Kevin Arceneaux, Alan S. Gerber, and Donald P. Green, *Political Analysis* 14: 1-36.

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

> The non-partisan civic group Vote 2002 Campaign ran a get-out-the-vote initiative to encourage voting in that year's U.S. congressional elections. In the 7 days preceding the election, Vote 2002 placed 60,000 phone calls to potential voters, encouraging them to "come out and vote" on election day.
>
> Did the program work? How can we estimate its impact?

Voter turnout has been in decline since the 1960s

While voter turnout (the number of eligible voters that participate in an election) has been declining since the 1960s, it was particularly low in the 1998 and 2000 U.S. elections. Only 47 percent of eligible voters voted in the 2000 congressional and presidential elections; the record low was 35 percent in the 1998 mid-term elections.

Vote 2002 get-out-the-vote Campaign

Facing the 2002 midterm election and fearing another low turnout, civic groups in Iowa and Michigan launched the Vote 2002 Campaign to boost voter turnout. The campaign employed telemarketing techniques commonly used in modern elections. In the week preceding the election, Vote 2002 placed phone calls to 60,000 voters and gave them the following message:

> *Hello, may I speak with [Mrs. Ida Cook] please? Hi. This is [Carmen Campbell] calling from* Vote 2002*, a non-partisan effort working to encourage citizens to vote. We just wanted to remind you that elections are being held this Tuesday. The success of our democracy depends on whether we exercise our right to vote or not, so we hope you'll come out and vote this Tuesday. Can I count on you to vote next Tuesday?*

As telephone campaigns replace many of the more traditional face-to-face interventions, there is considerable debate over their effectiveness. Many believe the decline in voter turnout is directly related to the reduction in more personal methods of campaigning. It is therefore worth asking in this context, did the Vote 2002 Campaign work? Did it increase voter turnout at the 2002 congressional elections?

Did the Vote 2002 Campaign work?

What is required in order for us to measure whether a program worked, 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. To say, validly, that a program changes outcomes, we need to establish three things: (1) that outcomes have changed; (2) that the observed changes occurred among participants of the program and did not occur among non-participants; and (3) that it is not something else, some other event happening at the same time as the program, that drove the observed changes. In other words, we need to show that the program *causes* the observed changes.

To show that the program causes the changes, we need to simultaneously show that if the program had not been implemented, the observed changes would not have happened. What is called the "counterfactual" is the imaginary state of the world that program participants would have experienced if they had not participated in the program. It does not represent the state in which would-be participants receive absolutely no services, but rather the state of the world in which life goes on as before, the participants receive whatever services they would have received had they not participated in the program; it represents life without participating in the program.

The impact of the program, then, is the difference between the observed outcomes and what those outcomes would have been in the absence of the program, under the counterfactual. Thus we need to know the counterfactual to determine impact. But the fact is the program was implemented; we can never observe the counterfactual. Because we cannot directly observe the true counterfactual, we cannot actually determine impact. The best we can do is to estimate it, and we do so by *mimicking* the counterfactual.

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—condition (3) above—we want to be able to say 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 Vote 2002 Campaign work? What was its impact on voter turnout?

In this case, the targeted behavior is to "get out and vote," and the outcome measure is voter turnout. So, when we ask if the Vote 2002 Campaign worked, we are asking if it increased voter turnout in the 2002 congressional elections. The impact is the difference between voter turnout on that Tuesday in 2002 and what voter turnout would have been if Vote 2002 had never existed.

What comparison groups can we use?

Estimating the impact of the Vote 2002 Campaign

Your team is doing pro-bono consulting for Vote 2002. Your task is to estimate the impact of the Vote 2002 Campaign. Vote 2002 had access to a list of the telephone numbers of 60,000 people. They called all 60,000, but they were able to speak to only 25,000. For each call, they recorded whether or not the call was completed successfully. They also had census data on the voter's age, gender, household size, whether the voter was newly registered, which state and district the voter was from and data on how competitive the previous election was in that district, and whether the individual had voted in the past. Afterwards, from official voting records, they were able to determine whether, in the end, the voters they had called did actually go out and vote.

There are a number of methods available to your team to estimate the impact. In this case, we will compare their validity and identify the circumstances under which a given method can be used or not.

Method 1: Using a simple difference

Discussion Topic 1: Using simple differences: comparing voter turnout between the "reached" and "not reached"

Method 1: Comparing voter turnout between reached and not reached. Assume the 25,000 households who received the full message constitute the participant group and the 35,000 households who were called but not reached represent the comparison group. If you want to see what the impact of receiving a call has on voter turnout, you could check whether those who were reached were more likely to vote than those who were not reached. Estimate impact by comparing the proportion of people who voted in the treatment group and that of the comparison group, as shown in the following table:

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

Discuss whether this method gives you an accurate estimate of the effect of the program. What might be the possible sources of biases? In other words, what is likely to make the comparison group a poor approximation of the true counterfactual?

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

Method 2: Using multivariate regression to control for inherent differences

Discussion Topic 2: Using multivariate regression

You were concerned that people reached might have different inherent characteristics from those who were not reached. Indeed, when you compare the two groups, you observe significant differences:

Characteristics of Reached and Not-Reached Groups

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

1. Can you overcome the problems of method 1 by taking a random sample from the participant group and a random sample from the comparison group?

Method 2: Using multivariate regression to control for differences between reached and not-reached.

You could control for these differences by using a multivariate regression as follows: The participant and comparison groups are defined in the same way as in method 1. To estimate the impact of the program, you run a regression where the "dependent variable" is a zero/one variable indicating whether the person voted or not (i.e., 0 = did not vote, 1 = voted). The "key explanatory variable" is a zero/one variable indicating whether the person received the call or not (i.e., 0 = did not receive the call, 1 = received a call). Potential differences in characteristics can be controlled for using other "explanatory variables" such as age, gender, newly registered voter, etc. The coefficient on the key explanatory variable (i.e., received the call) represents the "controlled" estimated impact of the program.

Using multivariate regression to control for the characteristics shown in the table below, you estimate the impact to be 6.1 pp (percentage points), significant at the 5% level.

2. Why do you think the estimated impact using method 2 is lower than the 10.8 pp impact you estimated using method 1?

3. For method 2, discuss whether it is reasonable to expect that the estimated impact represents the true causal effect of Vote 2002 on voter participation. What remaining biases could there be?

4. Using the data described above, can you think of more convincing methods to estimate the impact of the Vote 2002 Campaign?

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

Method 3: Using panel data—tracking the same people over time

You are still concerned about differences in characteristics between the reached and non-reached. You decide to use panel data, that is, track the same person over time.

Discussion Topic 3: Using panel data

Method 3: Using panel data to track the same people over time. It turns out that staff members of Vote 2002 also had data on whether the person voted in the previous elections (1998 and 2000). Past voting behavior is thought to be a strong predictor of future voting behavior. The table below indicates past voting behavior for the group of people who were reached by the Vote 2002 Campaign and the group of people who were called but not reached.

Voter turnout in 1998 and 2000 elections between the reached and not-reached

	2002 Reached	***2002 Not Reached***	***Difference***
Voted in 2000	71.7%	63.3%	8.3 pp*
Voted in 1998	46.6%	37.6%	9.0 pp*

1. How can these data on past voting behavior be used to improve your analysis?

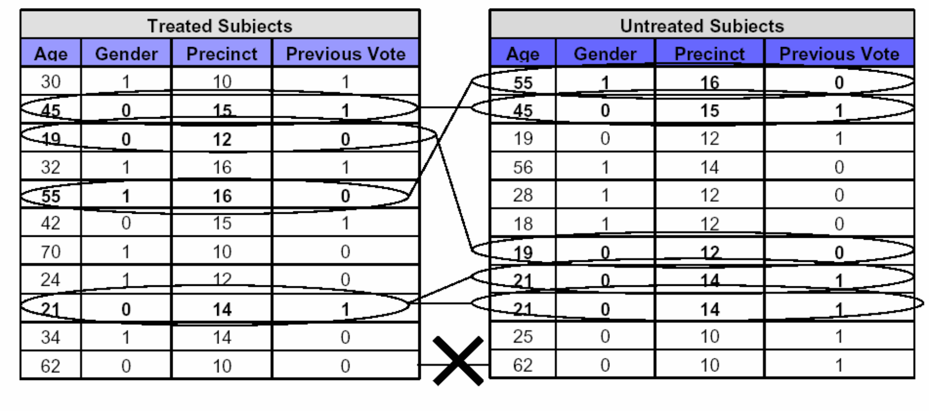
2. Given the information in the above table, would you expect that controlling for past voting behavior in method 2 would result in a higher or lower estimate of the impact of the Vote 2002 Campaign on voter turnout than the 6.1 pp found without controlling for it?

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

Method 4: Using matching

One way to estimate the impact of the Vote 2002 Campaign is to select as a comparison group a subset of non-participants who look similar to the participant group (the 25,000 called and reached). To select this subset, researchers often employ a statistical procedure called *matching*. While there are many ways to do matching, it turns out that in this context it is possible to do *exact matching* for almost all the individuals in the sample. The lists from which these 60,000 individuals were selected and tracked include data on another 2 million eligible voters. Therefore, for each of the 25,000 individuals reached, we can select another individual who has the exact same characteristics (i.e., age, gender, etc.). In this way, the participant and comparison groups will have exactly the same observable characteristics. Figure 1 shows exact matching.

Figure 1: Exact Matching



Treated Subjects

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

Untreated Subjects

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

Source: Arceneaux, Gerber, and Green (2004)

Discussion Topic 4: Exact Matching

Method 4: Matching. Matching was performed and then the impact of the Vote 2002 Campaign was estimated by taking the difference between the voter turnout rate in the participant group and the voter turnout rate in the comparison group created through matching (the "matched" group). The results are shown in the table.

Matching Analysis

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

1. Assess whether it is reasonable to expect that the impact estimated using this method represents the true causal effect of Vote 2002 on voter participation.

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

a. All: household size, age, newly registered, county, state senate district, state house district, from a competitive district, voted in 2000, voted in 1998. Using all covariates, only 90% of the reached-individuals had exact matches in the comparison group.

Method 5: Using randomized experiments

It turns out that from the larger population of about 2 million potential voters, the 60,000 individuals were ***randomly*** selected. Under the final method, the group that was called (whether reached or not reached) is now called the treatment group and the rest is the comparison group.

Discussion Topic 5: Randomized Experiment

Method 5: Randomized Experiment. You can exploit this randomization to estimate the impact of the Vote 2002 Campaign. The idea is that the 60,000 individuals Vote 2002 called (now called the treatment group) should be statistically identical to the 2,000,000 individuals (called the control group) in everything (observable and unobservable) except for the fact that the first group was called by the Vote 2002 Campaign.

Compares the treatment and control groups on observable characteristics

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

1. Notice that the two groups look very similar. Is this what you would expect?

Comparing voter turnout in the experimental treatment and the control groups

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

2. Notice that the impact estimates are not statistically significant. This result is different than those obtained with the previous methods. How do you explain this difference in results?

3. In the above analysis, we compare the 60,000 *who were called* to the 2,000,000 not called by the Vote 2002 Campaign. Why don’t we compare just the 25,000 who were *reached* to the same control group?

Adjusting estimate to remove “dilution” of impact from those not reached

	Impact
Difference after adjusting for the fact that only 25,000 of 60,000 in the treatment group were reached (“Treatment Effect on the Treated”)*	0.4 pp

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

* This corresponds to an instrumental variable regression that estimates the effect of the treatment “on the treated.”

Comparing all five methods

Below are the impact estimates of the Vote 2002 Campaign using the five different methods you have discussed in this case study.

Table 1: Comparing all five methods

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

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

As you can see, not all methods give the same result. Hence, the choice of the appropriate method is crucial. The purpose of this case study was not to evaluate one particular voter mobilization campaign, but to evaluate evaluation methods in this particular context.

In the analysis of the Vote 2002 Campaign, we found that people who happened to pick up the phone were more likely to vote in the upcoming (and previous) elections. Even though we statistically accounted for some observable characteristics, including demographics and past voting behavior, there were still some inherent, unobservable differences between the two groups, independent of the get-out-the-vote campaign. Therefore, when our non-randomized methods demonstrated a positive, significant impact, this result was due to "selection bias" (in this case, selection of those who pick up the phone) rather than a successful get-out-the-vote campaign.

Discussion Topic 6: Selection bias

Selection bias is a problem that arises in many program evaluations. Think about some of the non-randomized development programs you have, or have seen, evaluated. Discuss how the participant group was selected, and how "selection" may have affected the ability to estimate the true impact of the program.

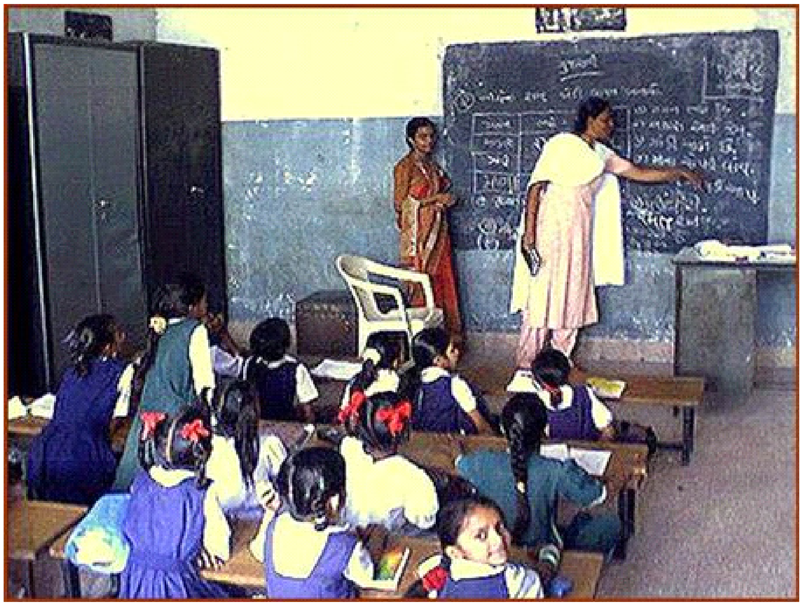
References:

Gerber, Alan and Donald Green, 2000. "The Effects of Canvassing, Telephone calls, and Direct mail on Voter Turnout: A Field Experiment" *American Political Science Review* 94 (3): 653-663

Arceneaux, Kevin, Alan Gerber, and Donald Green 2004. "Comparing Experimental and Matching Methods using a Large-Scale Field Experiment on Voter Mobilization" *Preliminary Draft*

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Massachusetts Institute of Technology



Case 2: Remedial Education in India

Evaluating the Balsakhi Program

Incorporating random assignment into the program

This case study is based on "Remedying Education: Evidence from Two Randomized Experiments in India," by Abhijit Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden (2007), ***Quarterly Journal of Economics*****, 122(3):1235-1264, 2007**

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

In 2000 the NGO Pratham was expanding its Balsakhi Program, a remedial education initiative, to 123 municipal primary schools in the city of Vadodara in western India. The program had been running in Mumbai since 1994, and Pratham wanted to take advantage of the expansion to conduct a randomized impact evaluation. The need for remedial education was general in the 123 Vadodara schools and, after an initial pilot, Pratham had enough resources to expand the program to all schools immediately, so there was a general sense that all eligible schools should receive program assistance. But how could Pratham have the program in all schools and, at the same time, keep the comparison group it needed for a randomized impact evaluation? How could random assignment be integrated into the program?

Children are in school but not learning

India has made much progress toward the Millennium Development Goal of universal primary education by 2015. Access to primary school has expanded, and more and more children are now participating: Net primary enrollment in 2005 was 89 percent. For many children, however, being enrolled does not necessarily mean learning much because the quality of schooling is often too low.

There are many reasons for low school quality.

Schools do not have enough resources and often have inappropriate curricula. There are too few teachers and some are poorly trained. There are also too few classrooms, teaching materials, textbooks, notebooks, and pencils. The curricula are often not adapted to the lack of resources or to the local context. Schools, therefore, fail to give basic academic education and the skills and knowledge students ultimately need to navigate their particular environment.

Teachers are often absent or make little effort when present. A countrywide survey found that one quarter of all public primary school teachers were absent from school on any given day and that only half of those present were teaching. [1]

Class size is often large. As more children enroll, pupil-teacher ratios worsen and teachers cannot give extra attention to pupils who may need it to follow the lesson. What's more, when the class size is larger, more of the teacher's attention has to be spent on ancillary classroom issues, such as discipline and simply getting the pupils coordinated and focused.

Not only are the classes large, but they also often include students of varying achievement or even grade levels. This makes it even more difficult to adapt the material and the pace to the learning needs of the pupils. The less-prepared pupils may need different instruction or a slower pace or even remedial education. But if the teacher focuses on their needs, the more-prepared students would be learning less.

Low school quality often translates into poor learning. In Mumbai, 25 percent of children in grades 3 and 4 in public schools cannot recognize letters, and 35 percent cannot recognize basic numbers; in Vadodara, only 19.5 percent of grade 3 students can correctly answer questions testing grade 1 math competencies. And a nationwide survey found that 44 percent of the in-school children aged 7 to 12 cannot read a basic paragraph and 50 percent cannot do simple subtraction.

[1] Part of a multinational survey that included 6 countries in different regions: Bangladesh, Ecuador, India, Indonesia, Peru, and Uganda.

Schools are failing to ensure that children are actually learning. Many who fall behind are promoted to upper grades before they have mastered the lower grade skills. Unprepared, they cannot follow the lessons and fall behind even further. Improving general school quality may not necessarily help these children if they don't have the basic skills they need to profit from the improvements—having your own grade 4 math textbook is little help if you can't do grade 1 math. But targeted initiatives that increase the basic skills children need to learn effectively could ensure that all children in school are also learning.

The Balsakhi Program provided remedial education

Pratham is an educational organization based in Mumbai whose motto is "Every child in school…and learning well."

In 1994 Pratham launched the Balsakhi Program to help at-risk children acquire the basic skills they need to participate fully in the classroom. The program provided tutors for at-risk children in government schools. The tutor, called a balsakhi, or "child's friend," was typically a young woman hired from the local community. Balsakhis were paid between 500 and 750 rupees (US$10-15) a month. All the balsakhis had completed at least secondary school, and they were given two weeks' training at the beginning of the school year.

The program targeted children who had reached grades 3 and 4 without mastering grades 1 and 2 reading and math competencies, including spelling simple words, reading simple paragraphs, recognizing numbers, counting up to 20, and subtracting or adding single-digit numbers. Children who were lagging behind—identified as such by the teacher—were pulled out of the regular class in groups of 20 and sent for remedial tutoring, spending half the school day with the tutor.

Tutoring followed a curriculum designed by Pratham to help the children acquire the grades 1 and 2 skills they needed to follow their regular lessons. But because the 20 pupils are pulled out of the regular classroom, the program could have two other potential effects. Pulling out the children created two classes, each smaller than the original. So for half of the school day, the class size was reduced. Pulling out the *weakest* children created two streams, each with children of comparable achievement. This amounted to tracking: For half the school day, a child in the regular class (the higher-ability track) temporarily had peers at an equal or more advanced learning level.

Therefore the impact of the program, if any, could come through one or more of the following channels: the remedial instruction delivered by the balsakhi, the reduction in class size, and the ability tracking.

Evaluation questions and designs

The opportunity to evaluate came when Pratham was expanding to Vadodara in 2000, six years after the program was launched in Mumbai. The objective of the program was to improve basic math and reading competence. In particular, Pratham wanted to make sure the program led to improvements in basic number recognition, counting, ordering one- and two-digit numbers, and solving basic word problems. Pratham also wanted to learn as much as possible about the channels through which the program achieves its impact.

Your team is invited to the very first evaluation planning session. The objective of the session is to decide on possible evaluation questions and corresponding designs. It has not emerged yet that all schools must get program assistance, so you can have some schools that do not receive balsakhis. Your task is to determine what you can learn from the different possible evaluation designs.

Discussion Topic 1: Possible evaluation questions and designs

For each of the following designs, say what comparisons you can make and what you can learn from them for each of the channels through which the program could have an impact.

1. Randomize at the school level. Half the schools receive balsakhis, and half the schools do not receive balsakhis.

2. Randomize at grade (cohort) level. Half the schools receive balsakhis in grade 3, and half the schools receive balsakhis in grade 4.

3. Randomize at individual level. Identify the weak students, and randomly select half of them to go to the balsakhi for half a day while the remaining weak students remain in the regular class.

Designing the evaluation considering the opportunities and the constraints

The pilot had shown that the need for remedial education was general in the municipal schools, so a general consensus emerged among the stakeholders that all schools had to receive balsakhis during the evaluation period. The decision to take part in the evaluation had been left to the schools. There was also some concern that schools would only be willing to take part in the evaluation—for example, allow Pratham to conduct achievement tests in the school—if they received some program assistance.

Whatever evaluation design was adopted, it had to ensure that all schools in the sample received the program and that somehow half the sample would be a comparison group not receiving the program.

Discussion Topic 2: Designing the evaluation to take advantage of the opportunities and resolve the most constraints

1. A crucial step in designing a randomized evaluation is to decide on the level to randomize. Choosing a particular level not only resolves constraints, it can also make the difference in what we can learn from an evaluation. This intervention is school-based, so you can randomize at the individual student, grade (cohort), or school level.
a. At what level is the program targeted?
b. What are the advantages and disadvantages of each of the possible levels?

2. Each of the following questions may represent a constraint you will face when deciding on the level of randomization.
For each possible level of randomization, discuss the following:
a. Are there potential spillovers: does providing balsakhis potentially affect those who are not treated?
b. Would randomizing at this level compromise the ethical, political, and practical feasibility?
c. Would there be enough units at this level for the design to have statistical power?

3. Pratham was particularly interested in learning the overall effects of the program on children in grades 3 and 4. Given the constraints and knowing what Pratham wanted to learn, at what level would you randomize?

4. If Pratham wanted to learn about the effects on the children sent to the balsakhi, what groups would you compare?

5. If Pratham wanted to learn about the effects on the children that remain in the regular class, what groups would you compare?

6. Synthesize your answers into a randomized design that you would use to take advantage of the opportunities and resolve the most constraints.
Create a chart that shows your randomization design and evaluation strategy.

The mechanics of simple random assignment

Now that you have a randomized evaluation design, you must do the actual random assignment. You need to have a list of units in your sample and the number of groups you will be assigning them to before you start. Once you have that, follow the procedure below to do a simple random assignment:

Step 1: Determine your allocation fraction. This is the proportion of units you will be assigning to the treatment group. The allocation fraction partly depends on your budget constraint.

Step 2: Order your sample randomly. Ordering the list randomly ensures that the position a school takes on the list is completely independent of any of its characteristics. This can be very easily done using a computer.

Step 3: Choose units from the randomly ordered list according to your allocation fraction. For example, if your allocation fraction is one-half you can take the top half of the entries or the bottom half and assign them to treatment.

Step 4: Check if your groups are equivalent for documentation purposes. If you have baseline data you can check if the groups are balanced on important characteristics. This involves comparing the averages of these characteristics across the groups. The test has to be based on data collected before the evaluation. And the characteristics have to be potential confounding factors, either observable intrinsic differences (gender, caste) or initial outcomes (income, level of education) that you think may factor into the final outcomes.

Discussion Topic 3: Simple random assignment

In Vadodara, the program was extended to 123 schools. Schools varied by language of instruction (Gujarati, Marathi, and Hindi), and by gender (schools were boys-only, girls-only, and co-educational). The evaluation was over two years. The problem was to ensure that all schools got balsakhis while also keeping a comparison group. So, every school got a balsakhi but for a specific cohort. Half the schools got balsakhis for grade 3 and half for grade 4. So, there were two groups. In Year 1, Group A got balsakhis for grade 3, and Group B got balsakhi for grade 4. In the following year, the schools switched; Group A got balsakhis for grade 4 and Group B for grade 3. That way, Group A children received the program for two years. Random assignment determined which schools got balsakhi for grade 3 or for grade 4 in the first year.

1. Use procedure outlined in Excel Exercise 2A and the data provided to randomly order the schools.

Can you predict any of the school's characteristics—for example, the area it is located in—based on its position in the sorted list?

2. Given the outcome of interest, what characteristics would you use to check the randomization? In what ways could these characteristics be confounding?

3. Are the groups balanced on these characteristics?

Some of the schools are boys-only (labeled "kumar"), some girls-only (labeled "kanya"), and some co-ed (labeled "mishra"). You want to ensure the treatment and comparison groups have the same proportions of boys' and girls' schools as the sample.

4. What would you change about simple random assignment to get a procedure that—for certain—yields groups are balanced?

The mechanics of stratified random assignment

With stratification, you first divide the sample into subgroups, or strata. All that is meant by "stratified random assignment" or "block random assignment" is that the sample was first divided into identifiable subgroups and then units were assigned randomly from those subgroups.

Besides balancing the groups by potential confounding factors, you may also want to stratify if you want to learn about program effects on particular subgroups, such as ethnic minorities, and there are very few in your sample. To ensure that there are some minorities in both the treatment and control groups, you should stratify. Stratification may also help with statistical power.

Here is the procedure for stratified random sampling.

Step 1: Divide the list into subgroups, or strata.
Step 2: Determine your allocation fractions for each subgroup, stratum.
Step 3: Order each list randomly.
Step 4: Choose from each list according to your allocation fraction.
Step 5: Document the averages for analysis.

In other words, divide the list into subgroups and then apply the simple random allocation procedure to each subgroup.

Discussion Topic 4: Stratified random assignment

1. Use procedure outlined above and the data provided in Excel Exercise 2B to do a stratified random assignment of the schools. Choose the characteristic you want to stratify on.

2. Are the groups balanced on these characteristics?

References:

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden (2007), "Remedying Education: Evidence from Two Randomized Experiments in India," *Quarterly Journal of Economics*, 122(3):1235-1264

Pratham Resource Center (2004), "Rapid Assessment of Learning Outcomes: All India, June-August 2004," www.pratham.org.

The PROBE Team (1999), Public Report on Basic Education in India, Oxford: Oxford University Press.

UNESCO, EFA Global Monitoring Report Team (2008). *Education for All by 2015: Will we make it?* Oxford:UNESCO Publishing, Oxford University Press

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Massachusetts Institute of Technology

Case 3: Women as Policymakers

Measuring the effects of political reservations
Thinking about measurement and outcomes

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

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

India amended its federal constitution in 1992, devolving power to plan and implement development programs from the states to rural councils, or Gram Panchayats (GPs). The GPs now choose what development programs to undertake and how much of the budget to invest in them. The states are also required to reserve a third of GP seats and GP chairperson positions for women. In most states, the schedule on which reserved seats and positions cycle among the GPs is determined randomly. This creates the opportunity to rigorously assess the impact of reservations 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. Their questions were what data to collect, what data collection instruments to use, and what sample size to plan for.

Empowering the Panchayati Raj

Panchayats have a long tradition in rural India. An assembly *(yat)* of five *(panch)* elders, chosen by the community, convened to mediate disputes between people or villages. In modern times Panchayats have been formalized into institutions of local self-government.

The impetus to formalize came from the independence leaders, who championed decentralized government. Gandhi favored village *(gram)* self-government *(swaraj)*, a system where every village would be "self-sustained and capable of managing its affairs." Prime-minister Nehru advocated giving the Panchayats "great power," so that villages would "have a greater measure of real *swaraj* in their own villages."

Thus Article 40 of the constitution India—adopted at independence—direct the states to ensure that the Panchayats "function as units of self-government." Implementation guidelines recommended a three-tier system, with village councils (*gram panchayat*) as the grassroot unit.[1] Most states followed both directive and guidelines so that by the early 1950s they had formalized Panchayats. But in the 1960s, with no real power and no political and financial support from the federal government, the Panchayats disappeared in most states. It was not until the 1990s that they were revived.

The revival came through the constitution. In 1992, India enacted the 73rd amendment, which directed the states to establish the three-tier Panchayati Raj system and to hold Panchayat elections every five years. Councilors are popularly elected to represent each ward. The councilors elect from among themselves a council chairperson called a pradhan. Decisions are made by a majority vote and the pradhan has no veto power. But as the only councilor with a full-time appointment, the pradhan 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 Panchayats. Funds still come from the central government but are no longer ear-marked for specific uses. Instead, the GP decides which programs to implement

[1] Village councils, called Gram Panchayats or GPs, 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.

and how much to invest in them. GPs can chose programs from 29 specified areas, including welfare services (for example, widows, care for the elderly, maternity care, antenatal care, and child health) and public works (for example, drinking water, roads, housing, community buildings, electricity, irrigation, and education).

Empowering women in the Panchayati Raj

The GPs are large and diverse. In West Bengal, for example, each has up to 12 villages and up 10,000 people, who can 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 policymakers' identities influence policy choices, then groups underrepresented in politics and government could be shut out as GPs could ignore those groups' policy priorities. There were fears that the newly empowered GPs would undermine the development priorities of traditionally marginalized groups—the scheduled castes (SCs), scheduled tribes (STs), and women. To forestall this, the 73rd amendment included two mandates to ensure that investments reflected the needs of everyone in the GP.

The first mandate secures community input. If GP investments are to reflect a community's priorities, the councilors must first know what those priorities are. So, GPs are required to hold a general assembly (*gram sabha*) 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 Pradhans 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 the SC, ST, and women. States are required to reserve seats and pradhan positions for SC and ST in proportion to their share of the population and to reserve at least a third of all council seats and pradhan positions for women. Furthermore, the states have to ensure that the seats reserved for women are "allotted by rotation to different constituencies in a Panchayat" and that the pradhan positions reserved for women are "allotted by rotation to different Panchayats." In other words, they have to ensure that reserved seats and pradhan positions rotate evenly within and among the GPs.

Reserved seats and positions are randomly allocated

In most states, the order of the rotation is determined randomly. Random allocation is based on a table of random numbers in the Panchayat Electoral Law. GPs are ranked in order of their legislative serial number, and the table is then used to determine the seats reserved for SCs and STs (it provides the rank of the GP to assign to each list). GPs are then placed in three separate lists, again ranked by their number: the first consists of GPs reserved for SCs, the second, GPs reserved for STs, and the last, unreserved GPs. Then, in the first election, every third GP in each list starting with the first is reserved for a woman pradhan. Thus, some villages are reserved for an SC woman, some for an ST woman, and some for a woman in general. In the second election, the process to create the SC and ST list is repeated (with a new set of ranks assigned to each list), and every third GP starting with the second on each list is reserved for a woman, and so on.

Randomized reservation in India: What can it teach us?

Your evaluation team has been entrusted with the opportunity to estimate the impact of reservations for women in the Panchayats. Your evaluation should address all dimensions in which reservations for women are changing local communities in India. What data will you collect? What instruments will you use? How large will your sample be?

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

Discussion Topic 1: Gender reservations in the Panchayati Raj

1. What were the main goals of the Panchayati Raj?

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 reservations for women?

Gender reservations 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, 17 of the top 20 countries in the rankings have reservations.

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

4. Debate the pros and cons of reservations. At the end of the debate, you should have a list of the pros and cons of reservations.

5. What evidence would you collect to strengthen the case of each party?

6. Both parties are concerned about the causal impact of reservations at the national level. They appoint a bipartisan "methodology" commission to agree on what type of research to accept as evidence. Your team is called to give expert testimony on what constitutes good evidence. The first question you are asked is whether it is at all possible to determine the causal impact of reservations at the national level.

What will you tell the commission?

7. A commissioner brings up randomized reservations in India and asks if there are differences between the situation in India and the situation the commission is reviewing.

What will you tell the commissioner?

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 reservation policy should be evaluated? In which areas do you expect to see a difference as a result of reservations?

What are all the possible effects of reservations?

Discussion Topic 2: Using a logical framework to delineate your intermediate and final outcomes of interest

1. Brainstorm the possible effects of reservations, both positive and negative.

Hint: Use your answers to Discussion Topic 1 as a starting point.

2. For each potential effect on your list, list also the indicator(s) you would use for that effect. For example, if you say that reservations will affect political participation of women, the indicator could be ‟number of women attending the Gram Sabha.”

3. Suppose you had all the money and resources in the world and could collect data on every one of these indicators in reserved and unreserved communities, and compare them. How many indicators would you collect?

Multiple outcomes are difficult to interpret, so define a hypothesis

Reservations for women could affect a large number of outcomes in different directions. For example, it 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 reservation policy, it will be very difficult to make sense of any result we find. Think of the following: if you took 500 villages, and randomly assigned them in your computer to a “treatment” group and a “control” group, and then run regressions to see whether the villages look different along 100 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 just present your report in front of the commission who mandated you to evaluate this policy, explaining that the reservation for women changed some variables and did not change others, what are they supposed to make of it? How will they know that these differences are not due to pure chance, rather than the policy? You need to present them with a clear hypothesis of how reservations are supposed to change policymaking, which will lead you to make predictions about which outcomes are affected.

Discussion Topic 2: Using a logical framework to delineate your intermediate and final outcomes of interest

4. What might be some example of key hypotheses you would test? Pick one.

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

6. What is the possible chain of outcomes in the case of reservations?

7. What are the main critical steps needed to obtain the final results? What are the conditions needed to be met at each step?

8. What variables should you try to obtain at every step in your logical framework?

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

What data collection instruments to use

Now that you have determined outcomes of interest, there are several methods available to your team to answer your questions about the effects of reservations. Here are some examples of what you could do, including their costs in man days.

Pradhan Interview. You can interview the pradhan. This can give you information on socioeconomic background, political ambitions, and investments made since taking office. It costs one man day per pradhan interview.

Participatory Resource Appraisal (PRA). This method involves drawing a map of the village with the help of 10 to 20 villagers. Figure 1 shows a map from Damdama Phanchayat. The map shows public infrastructure (schools, wells, roads; SC and ST areas; cultivated land and energy projects). You can also find out when the infrastructure was built or repaired, what its (perceived) quality is; and also about the participation of women in various activities. It costs three man days to complete a very detailed PRA and to complement it with focus groups. It will cost an additional man day to travel between GPs, and half a man day to travel between villages in a GP.

Transcript of the Gram Sabha. You can send members of your team to record the Gram Sabha. The transcript will give you information about who speaks (gender), when, how long, and what they speak about (water, schools, governance). Attending the meeting and transcribing and translating takes a long time. It costs at least five man days for each Gram Sabha covered.

Household Surveys. You can interview a sample of the households to obtain both objectives and perceptions from all household members. Along with the PRA the household surveys allow very detailed data to be collected. However, this is the most expensive method of data collection. First, you need to start with a PRA to establish the household list from which your household can be sampled. A short questionnaire (a simple questionnaire without physical measurement) and focusing on interviewing only one or two household members will cost the research group roughly half a man day per household. A long questionnaire (involving health measurement for example) will cost up to a full man day per household.

Existing Administrative Data. You can also ask the pradhan for the GP balance sheets, which are supposed to be public information. You can also obtain minutes of past Gram Sabhas as part of the village PRA and the latest national census data. Data from the 1991 and 2001 censuses are available. It cost zero man days to get access to census data.

Table 1 summarizes each of the methods available to your team.

Discussion Topic 3: Data collection instruments

1.	What are the advantages and disadvantages of each of the tools?
2.	If you had an unlimited budget, what tools would you use to collect your data?
3.	If you had a limited budget, what tools would you use to be able to test your hypothesis?
4.	What instruments would you use to collect data on policy preferences?

Figure 1: An actual PRA from Damdama Panchayat

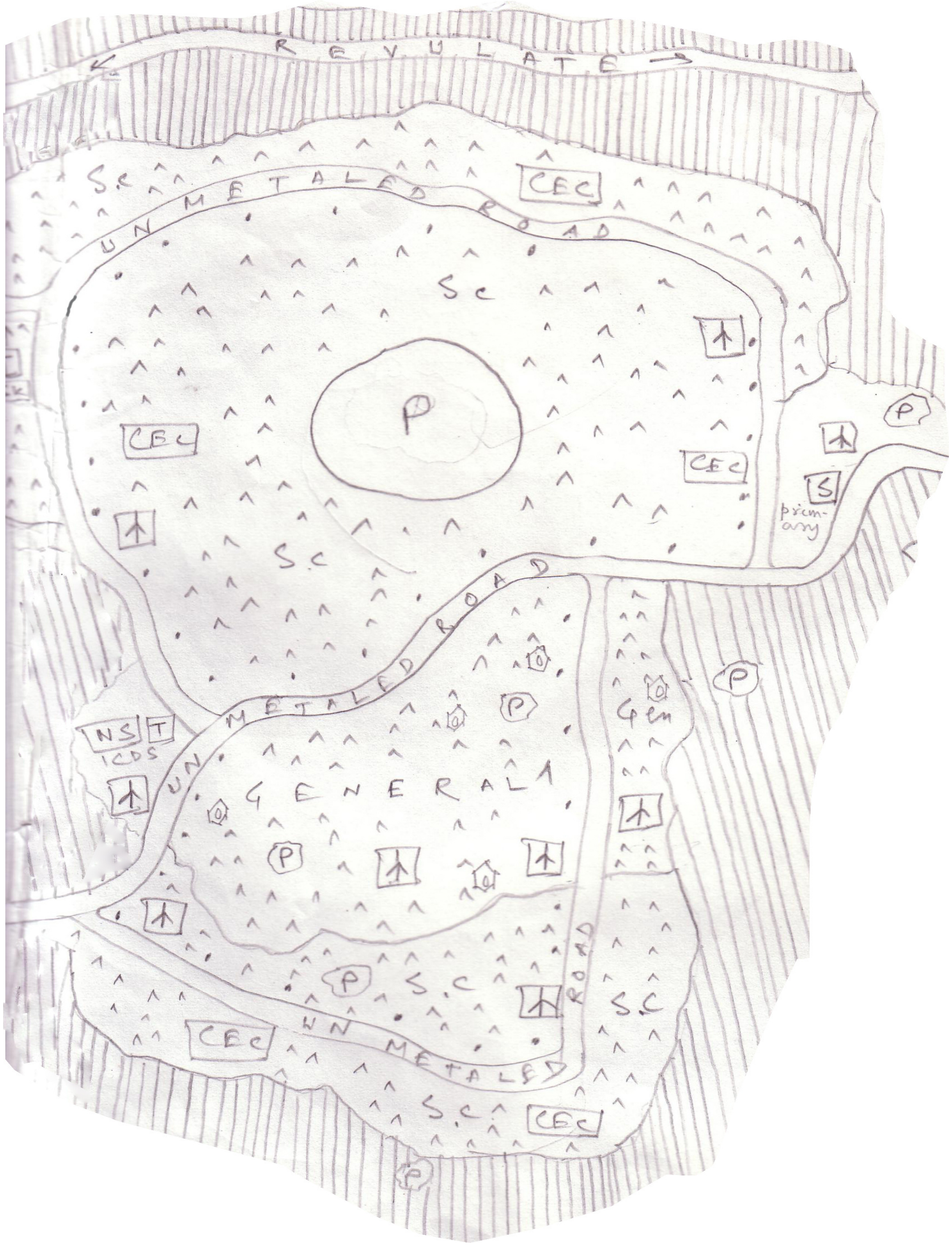


Table 1: Data collection instruments

Tool	*Target Respondent*	*Target Outcomes*	*Cost**
GP Interview	Pradhan	o Pradhan's background (socioeconomic status, education) o Political ambitions o Political experience o Investments undertaken o Public records. such as GP balance sheets	o Cost = 1 man day per interview o Travel Cost between GPs = 1 man day
Transcript of Gram Sabha	GP	o Who speaks and when (gender) o For how long do they speak? o What issues do they raise?	o Cost = 5 man days for every meeting attended, transcribed, and translated
Complaints and service requests	GP	o What have men and women complained about?	o Cost = 0 man days
Village Participatory Resource Appraisal (village mapping exercise and focus groups)	10 to 20 villagers per village	o Village infrastructure (schools, roads, wells, SC and ST areas, cultivated land, irrigation, energy projects) o Perception of quality of different public goods o Participation of men and women in activities o What issues villagers have raised with GP	o Cost = 3 man days for every map drawn and focus group conducted o Travel Cost between villages in GP = ½ man day
Household interviews	Head of household (the male in some HH; the female in other HH)	o HH demographic and socioeconomic data o HH outcomes (child heath, measurement of height and weight, etc.) o HH perceptions of quality of public goods and services o Declared HH preferences	o Short questionnaire with no physical measurement = ½ man day per HH; o Long questionnaire with physical measurement = 1 man day per HH
Existing administrative data	Public data archives (national, GP, and Village)	o A snapshot of village characteristics—population, public goods, demographics, etc.—at the time of the 1991 and 2001 census o Expenditures on public goods and services in GP (from GP balance sheets) o Issues addressed at GP public assemblies (from Gram Sabha minutes)	o Cost = 0 man days

* Costs are given in man days. We will assume here that all other expenses can be computed using a simple overhead rule. Anything with a cost of zero is charged to overhead.

How much data to collect—planning the sample size

To be able to draw credible conclusions about the general population, the sample of GPs you use for your study must be representative of the general population. How large does the sample size need to be to credibly detect an effect size? By credibly we mean only that you can be reasonably sure that the difference in outcomes you see between the reserved and unreserved GPs is due the reservation policy. Randomization removes bias, but it does not remove noise; it ensures comparability because of the law of large numbers. The question is, how large must large be? What sample size do you need to be able to test your hypotheses of interest?

Discussion Topic 4: Power and Cost Tradeoffs

Use Excel exercise 3B to answer the following questions:

1. How does power vary with sample size?

2. How does power vary with effect size?

3. How does power vary with the level of clustering?

4. What effect size do you think you need to be able to tell if women had an impact on investments in drinking water?

5. Given the effect size you chose in (4), what is the smallest number of villages you need to detect the effect?

6. Does the study with the smallest number of villages have the smallest budget? Why? Why not?

7. How should you pick the minimum number of villages to have the smallest budget to detect the effect?

8. How many villages do you need?

9. Given a budget of 900 man days, what data would you collect? Of the questions you are interested in, which could you answer? Which questions that you would want to answer are you unable to answer within this budget (these questions come from the budget for various things before)?

10. Given a budget of 4000 man days, what data would you collect? What new questions could you answer?

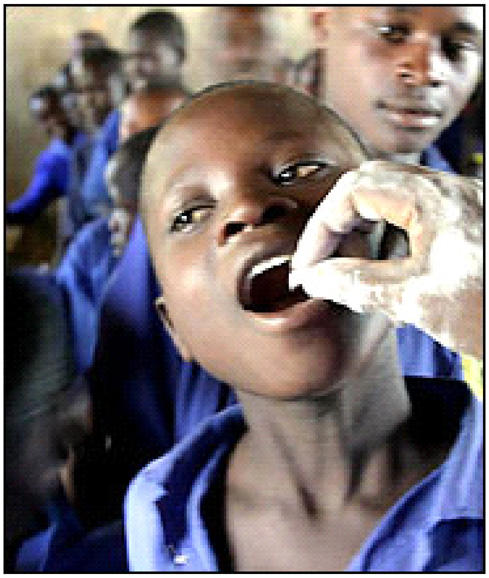
References:

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

UNICEF (2008). The State of the World's Children 2007: Women and Children, the double dividend of gender equality. New York, New York: UNICEF

ABDUL LATIF JAMEEL
Poverty Action Lab
TRANSLATING RESEARCH INTO ACTION

Massachusetts Institute of Technology



Case 4: Deworming in Kenya

Managing threats to experimental integrity

This case study is based on Edward Miguel and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," ***Econometrica*** **72(1): 159-217, 2004**

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

Between 1998 and 2001, the NGO International Child Support Africa implemented a school-based mass deworming program in 75 primary schools in western Kenya. The program treated the 30,000 pupils enrolled at these schools for worms—hookworm, roundworm, whipworm, and schistosomiasis. Schools were phased-in randomly.

Randomization ensures that the treatment and comparison groups are comparable at the beginning, but it cannot ensure that they remain comparable at the end of the program. Nor can it ensure that people comply with the treatment they were assigned. Life also goes on after the randomization: other events besides the program happen between randomization and the end-line. These events can reintroduce selection bias; they diminish the validity of the impact estimates and are threats to the integrity of the experiment.

How can common threats to experimental integrity be managed?

Worms—a common problem with a cheap solution

Worm infections account for over 40 percent of the global tropical disease burden. Infections are common in areas with poor sanitation. More than 2 billion people are affected. Children, still learning good sanitary habits, are particularly vulnerable: 400 million school-age children are chronically infected with intestinal worms.

Worms affect more than the health of children. Symptoms include listlessness, diarrhea, abdominal pain, and anemia. Beyond their effects on health and nutrition, heavy worm infections can impair children's physical and mental development and reduce their attendance and performance in school.

Poor sanitation and personal hygiene habits facilitate transmission. Infected people excrete worm eggs in their feces and urine. In areas with poor sanitation, the eggs contaminate the soil or water. Other people are infected when they ingest contaminated food or soil (hookworm, whipworm, and roundworm), or when hatched worm larvae penetrate their skin upon contact with contaminated soil (hookworm) or fresh water (schistosome). School-age children are more likely to spread worms because they have riskier hygiene practices (more likely to swim in contaminated water, more likely to not use the latrine, less likely to wash hands before eating). So treating a child not only reduces her own worm load, it may also reduce disease transmission—and so benefit the community at large.

Treatment kills worms in the body, but does not prevent re-infection. Oral medication that can kill 99 percent of worms in the body is available: albendazole or mebendazole for treating hookworm, roundworm, and whipworm infections; and praziquantel for treating schistosomiasis. These drugs are cheap and safe. A dose of albendazole or mebendazole costs less than 3 US cents while one dose of praziquantel costs less than 20 US cents. The drugs have very few and minor side effects.

Worms colonize the intestines and the urinary tract, but they do not reproduce in the body; their numbers build up only through repeated contact with contaminated soil or water. The WHO recommends presumptive school-based mass deworming in areas with high prevalence. Schools with hookworm, whipworm, and roundworm prevalence over 50 percent should be mass treated with albendazole every 6 months, and schools with schistosomiasis prevalence over 30 percent should be mass treated with praziquantel once a year.

Primary School Deworming Program

International Child Support Africa (ICS) implemented the Primary School Deworming Program (PSDP) in the Busia District in western Kenya, a densely-settled region with high worm prevalence. Treatment followed WHO guidelines. The medicine was administered by public health nurses from the Ministry of Health in the presence of health officers from ICS.

The PSDP was expected to affect health, nutrition, and education. To measure impact, ICS collected data on a series of outcomes: prevalence of worm infection, worm loads (severity of worm infection); self-reported illness; and school participation rates and test scores.

Evaluation design — the experiment as planned

Because of administrative and financial constraints the PSDP could not be implemented in all schools immediately. Instead, the 75 schools were randomly divided into 3 groups of 25 schools, and phased-in over 3 years. Group 1 schools were treated starting in both 1998 and 1999, Group 2 schools in 1999, and Group 3 starting in 2001. Group 1 schools were the treatment group in 1998, while schools Group 2 and Group 3 were the comparison. In 1999 Group 1 and Group 2 schools were the treatment and Group 3 schools the comparison.

Figure 1: The planned experiment: the PSDP treatment timeline showing experimental groups in 1998 and 1999

	1998	1999	2001
Group 1	**Treatment**	**Treatment**	**Treatment**
Group 2	Comparison	**Treatment**	**Treatment**
Group 3	Comparison	Comparison	**Treatment**

Threats to integrity of the planned experiment

Discussion Topic 1: Threats to experimental integrity

Randomization ensures that the groups are equivalent, and therefore comparable, at the beginning of program. The impact is then estimated as the difference in the average outcome of the treatment group and the average outcome of the comparison group. To be able to say that the program caused the impact, you need to be able to say that the program was the only difference between the treatment and comparison groups over the course of the evaluation.

1. What does it mean to say that the groups are equivalent at the start of the program?

2. Can you check if the groups are equivalent at the beginning of the program? How?

3. What can happen over the course of the evaluation to make the groups non-equivalent?

4. How does non-equivalence at the end threaten the integrity of the experiment?

5. You randomized, creating equivalent treatment and comparison groups. If the groups remain equivalent, what else can happen after randomization to threaten your ability to say the program was the only difference between the two groups?

6. In Case 1, you learned about other methods to estimate program impact, such as simple difference, multiple regression, multiple regression with panel data, and matching.

- **a.** For each threat you just identified, say if and how the threat exists for each of these methods.
- **b.** Are the threats to experimental integrity unique to randomization?

Managing attrition—when the groups do not remain equivalent

Attrition is when people join or drop out of the sample—both treatment and comparison groups—over the course of the experiment. One common example in clinical trials is when people die; so common indeed that attrition is sometimes called experimental mortality.

Discussion Topic 2: Managing Attrition

You are looking at the health effects of deworming. In particular you are looking at the worm load (severity of worm infection). Worm loads are scaled as follows: Heavy worm infections get a worm load score of 3, medium worm infections a score of 2, and light infections a score of 1.

The program is school-based, so it is natural and cost-effective to collect data at the schools—the children are gathered in one place, so the enumerator does not have to travel to every child's home. The enumerator takes the measurements on all children in school on a randomly chosen day (the school authorities are not given prior warning).

There are 30,000 children, 15,000 in treatment schools and 15,000 in comparison schools. After you randomize the groups are equivalent, children from each of the three categories are equally represented.

Protocol compliance is 100 percent: all children who are in the treatment get treated and none of the children in the comparison are treated. Deworming at the beginning of the school year results in a worm load of 1 at the end of the year because of re-infection. Children who have a worm load of 3 only attend half the time, and drop out of school if they are not treated. The number of children in each worm-load category is shown for both the pretest and posttest.

	Pretest		***Posttest***	
Worm Load	Treatment	Comparison	Treatment	Comparison
3	5,000	5,000	0	**Dropped out**
2	5,000	5,000	0	5,000
1	5,000	5,000	15,000	5,000
Total children tested at school	15,000	15,000	15,000	10,000

1. **a.** What is the average pretest worm load for the treatment group?
b. What is the average pretest worm load for the comparison group?
c. Are the groups equivalent?

2. **a.** What is the average posttest worm load for the treatment group?
b. What is the average posttest worm load for the comparison group?
c. What is the difference?

3. **a.** Calculate the outcome differences at the beginning and at the end of the year?
b. Is this outcome difference an accurate estimate of impact of the program?
c. If it is not accurate does it overestimate or underestimate the impact?

4. Because the treatment was treated, you expected there to be a difference between the groups at the end of the year.
a. If this difference is an effect, what is the source of attrition bias, if any?
b. How can you solve the problem to get a better estimate of program impact?

5. **a.** What is the average posttest worm load for the comparison group if you also tested the 5,000 dropouts (assuming all would have had worm loads of 3)?
b. Calculate the impact of the program.
c. What is the size of the attrition bias?

6. **a.** The PSPD also looked at school attendance rates and test scores.
b. Would differential attrition bias either of these outcomes?
c. Would the impact be underestimated or overestimated?

7. In their song *A Day in the Life*, the Beatles sing, "And though the holes were rather small, they had to count them all."

Why should your consider adopting *A Day in the Life* as your theme song when you are thinking about managing attrition?

Managing partial compliance—when the treatment does not actually get treated or the comparison gets treated

Some people assigned to the treatment may in the end not actually get treated. In an after-school tutoring program, for example, some children assigned to receive tutoring may simply not show up for tutoring. And the others assigned to the comparison may obtain access to the treatment, either from the program or from another provider. Or comparison-group children may get extra help from the teachers or acquire program materials and methods from their classmates. Either way, these people are not complying with their assignment in the planned experiment. This is called "partial compliance" or "diffusion" or, less benignly, "contamination." In contrast to carefully-controlled lab experiments, diffusion is ubiquitous in social programs. After all, life goes on, people will be people, and you have no control over what they decide to do over the course of the experiment. All you can do is plan your experiment and offer them treatments. How then can you manage threats arising from partial compliance?

Discussion Topic 3: Managing partial compliance

All the children from the poorest families don't have shoes and so they have worm loads of 3. Though their parents had not paid the school fees, the children were allowed to stay on in school during the year. Parental consent was required for treatment and to give consent, the parents had to come to the school and sign a consent form in the headmaster's office. Because they had not paid school fees, the poorest parents were reluctant to come to the school. So none of the children with worm loads of 3 were actually treated. Their worm loads scores remain 3 at the end of the year. No one assigned to comparison was treated. All the children in the sample at the beginning of the year were followed up, if not at school then at home.

	Pretest		***Posttest***	
Worm Load	Treatment	Comparison	Treatment	Comparison
3	5,000	5,000	5,000	5,000
2	5,000	5,000	0	5,000
1	5,000	5,000	10,000	5,000
Total children tested	15,000	15,000	15,000	15,000

1. **a.** Calculate the impact estimate based on the original assignments.
b. What does this "intention to treat" estimate measure?
c. This is an accurate measure of the effect of the program, but is it a good measure? What are the considerations? When is it useful? When is it not useful?

You are interested in learning the effect of treatment on those actually treated.

2. Five of your colleagues are passing by your desk; they all agree that you should calculate the effect of the treatment using only the 10,000 children who were treated.
a. What is the impact using only the treated?
b. Is the advice sound? Why? Why not?

3. Another colleague says that it's not a good idea to drop the untreated entirely; you should use them but consider them as part of the comparison.
a. What is the impact estimate based on this strategy?
b. Is the advice sound? Why? Why not?

4. Another colleague suggests that you use the compliance rates, the proportion of people in each group that complied with the treatment assignment. You should divide the "intention to treat" estimate with the difference in the compliance rates.
a. What are the compliance rates in the treatment and comparison groups?
b. What is the impact estimate based on this strategy?
c. Is the advice sound? Why? Why not?

5. The program raised awareness of the worms, so some parent in the comparison bought the drugs and treated the children at home. Altogether 2,000 comparison children were treated.

What is the "treatment on the treated" impact estimate?

Managing spillovers—when the comparison, itself untreated, benefits from the treatment being treated

People assigned to the control group may benefit indirectly from those receiving treatment. For example, a program that distributes insecticide-treated nets may reduce malaria transmission in the community, indirectly benefiting those who themselves do not sleep under a net. Such effects are called externalities or spillovers.

Discussion Topic 4: Managing spillovers

In the PSPD, randomization was at the school level.

People in the evaluation areas lived on farms close together. Clusters of farms can be divided into areas of 3km radius. Three such areas—A, B, and C—are shown in the diagram below.*Farms are closed enough for children from neighboring farms to play with one another. Families also had a choice of primary schools.

There are three schools in area A, three in area B, and five in area C. It was common for children from neighboring farms, or even siblings, to go to different schools. Some of the schools in each cluster were treatment, others were control. Group 1 schools were the treatment in year 1, and group 2 and 3 were the comparison.

Each school has 100 children. Protocol compliance is 100 percent: all the children in treatment get treated and all the children in comparison do not get treated.

1. You estimate impact by comparing average worm loads at treatment and comparison schools.

Would this estimate be an underestimate or overestimate of the impact?

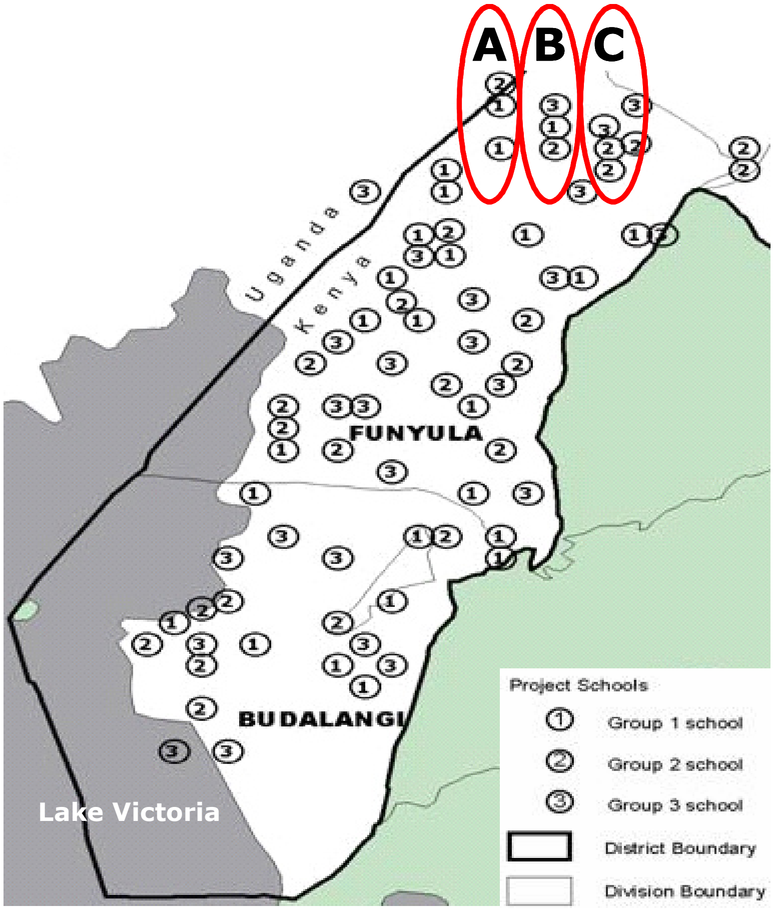
2. **a.** The treatment density is the proportion of treated to untreated in a given grouping of people.
b. What is the treatment density at the treatment schools in year 1?
c. What is the treatment density of comparison schools?
d. What are the treatment densities in areas A, B, and C in year 1?
e. What are the treatment densities in areas A, B, and C in year 2 and year 3?

3. **a.** If there are any spillovers, where would you expect them to come from?
b. Is it possible for you to capture spillover effects within the schools?
c. If you don’t expect to be able to capture the spillover effect, what would you need to be able to capture them?
d. Is it possible for you capture cross-school spillovers?

4. Rank the areas A, B, and C in terms of the amount of treatment spillover effects expected in years 1, 2, and 3.

5. **a.** If you had randomized at the individual level, what could you have done to capture interpersonal spillover?
b. If you had randomized at the school level what can you do to capture cross-school spillovers?
c. What general strategy does this suggest?

Discussion Topic 4: Managing spillovers



[*] The GPS locations were collected before May 2000, when the U.S. was still downgrading international GPS accuracy. Readings may only be accurate to within several hundred meters. So one Group 3 school appears to be in Uganda, but it's actually on the Kenyan side of the border. The school that appears to be in Lake Victoria is actually on a very small island.

References:

Crompton, D.W.T. 1999. "How Much Helminthiasis Is There in the World?" Journal of Parasitology 85: 397 – 403.

Kremer, Michael and Edward Miguel. 2007. "The Illusion of Sustainability," Quarterly Journal of Economics 122(3)

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," Econometrica 72(1): 159-217.

Shadish, William R, Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin Company

World Bank. 2003. "School Deworming at a Glance," Public Health at a Glance Series. http://www.worldbank.org/hnp

WHO. 1999. "The World Health Report 1999," World Health Organization, Geneva.

WHO. 2004. "Action Against Worms" Partners for Parasite Control Newsletter, Issue #1, January 2004, www.who.int/wormcontrol/en/action_against_worms.pdf

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 Project Sponsored by
THE COUNCIL FOR Excellence IN GOVERNMENT

August 2007

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@excelgov.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-designed randomized controlled trials are needed to produce strong evidence that an intervention is effective.

Checklist for overall study design

□ **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])

□ **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-designed 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

□ **The study report includes an analysis showing there are few or no systematic differences between the intervention and control groups prior to the intervention (e.g., in age, sex, income, education).**

□ **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).**

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

□ **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.

□ **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

□ **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]

□ **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 fluency, and not just the ability to sound out words.

□ **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.

□ **The study preferably obtained data on long-term outcomes of the intervention (e.g., a year after the intervention ended, preferably longer).**

This enables policymakers and practitioners to judge whether the intervention's effects were sustained over time. In most cases, it is the longer-term effects, rather than the immediate effects, that are of greatest policy and practical importance.

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

□ **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.

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-designed 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-designed 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://www.whatworkshelpdesk.ed.gov/guide_RCT.pdf, 2005.

[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 Bloom, *Randomizing Groups to Evaluate Place-Based Programs* (http://www.wtgrantfoundation.org/usr_doc/RSChapter4Final.pdf), prepared for a conference of the Society for Research on Adolescence, March 2, 2004.

[5] Here are illustrative examples of sample sizes from well-designed 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/Default.aspx?tabid=157; (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/Default.aspx?tabid=35; 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/Default.aspx?tabid=163; 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/Default.aspx?tabid=116.

[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.