

ABDUL LATIF JAMEEL
Poverty Action Lab



TRANSLATING RESEARCH INTO ACTION

Evaluating Social Programs

May 24–28, 2010

Executive Education at the Jameel Poverty Action Lab

Evaluating Social Programs Course Schedule 2009

May 24, 2010 – May 28, 2010

MIT, Room E51-395

Monday, May 24

8:00 AM – 9:00 AM	Continental Breakfast
9:00 AM – 9:15 AM	Opening Remarks
9:15 AM – 10:30 AM	<u>Lecture 1: What is Evaluation?</u> Lecturer: Rachel Glennerster (MIT)
10:30 AM – 12:00 PM	Case 1 (Women as Policy Makers): Group Discussion and exercise
12:00 PM – 1:30 PM	Lunch
1:30 PM – 3:00 PM	<u>Lecture 2: Outcomes, Indicators, and Measuring Impact</u> Lecturer: Marc Shotland (MIT)
3:30 PM – 6:00 PM	Group Project Work (choose topics for presentation)

Tuesday, May 25

8:00 AM – 8:30 AM	Continental Breakfast
8:30 AM – 10:00 AM	Case 2 (Learn to Read Evaluations): Group Discussion and exercise
10:30 AM – 12:00 PM	<u>Lecture 3: Impact Evaluation – Why Randomize?</u> Lecturer: Dan Levy (Harvard University)
12:00 PM – 1:30 PM	Lunch
1:30 PM – 3:00 PM	<u>Lecture 4: How to Randomize?</u> Lecturer: Leigh Linden (Columbia University)
3:30 PM – 4:45 PM	Exercise 1: Mechanics of Randomization
3:30 PM – 5:00 PM	Case 3 (Extra Teacher Program): Group Discussion and exercise
5:00 PM – 5:30 PM	Group Project
5:45 PM- 8:30 PM	Dinner at “The Elephant Walk” http://www.elephantwalk.com/

Wednesday, May 26

8:00 AM – 8:30 AM	Continental Breakfast
8:30 AM – 9:15 AM	Exercise 2: Random Sampling and Law of Large Numbers
9:30 AM – 12:00 PM	<u>Lecture 5: Sampling and Sample Size</u> Lecturer: Ben Olken (MIT)
12:00 PM – 1:30 PM	Lunch
1:30 PM – 3:00 PM	Exercise 3: Power Calculations, Sample Size
3:00 PM – 4:30 PM	<u>Lecture 6: Implementing an Evaluation</u> Lecturer: Shawn Cole (Harvard University)
5:00 PM – 6:00 PM	Group Project

Thursday, May 27

8:00 AM – 8:30 AM	Continental Breakfast
8:30 AM – 10:00 AM	Case 4 (Deworming in Kenya): Group Discussion and Exercise
10:30 AM – 12:00 PM	<u>Lecture 7: Analysis and Inference</u> Lecturer: Shawn Cole (Harvard University)
12:00 PM – 1:30 PM	Lunch
1:30 PM – 3:00 PM	<u>Lecture 8: Randomized Evaluation: Start-to-Finish</u> Lecturer: Nava Ashraf (Harvard University)
3:30 PM – 6:00 PM	Group Project: finalize presentation

Friday, May 28

8:00 AM – 9:00 AM	Continental Breakfast
9:00 AM – 12:00 PM	Group Presentations
12:00 PM – 1:00 PM	Lunch
1:00 PM – 3:00 PM	Group Presentations
3:00 PM – 4:00 PM	Course Wrap Up



Lecturer Bios

Nava Ashraf

Nava Ashraf is an Assistant Professor at Harvard Business School. Her research focuses on how people make decisions, applying principles from economics and psychology to design more effective development interventions. She has conducted randomized evaluations of savings innovations in the Philippines, an agricultural marketing intervention in Kenya, and is currently working on a randomized evaluation of socially-marketed health products in Zambia. She received her Ph.D in Economics from Harvard University in 2005 and her B.A. in Economics and International Relations for Stanford University in 1998.

Shawn Cole

Shawn Cole is an assistant professor in the Finance Unit at Harvard Business School, where he currently teaches the first half of the required finance course in the MBA program. His research examines corporate finance and banking in developing countries, covering topics such as bank competition, government regulation, and how financial development affects economic growth. Before joining the Harvard Business School, Professor Cole worked as an assistant economist at the Federal Reserve Bank of New York. He served as chair of the endowment management committee of the Telluride Association, a non-profit educational organization, for several years. He received a Ph.D. in Economics from the Massachusetts Institute of Technology in 2005, where he was an NSF and Javits Fellow, and an A.B. in Economics and German Literature from Cornell University.

Rachel Glennerster

Rachel Glennerster joined the Abdul Latif Jameel Poverty Action Lab at MIT as Executive Director in 2004. She earned her B.A. in Philosophy, Politics, and Economics from Oxford University and her Ph.D. in Economics from the University of London. She was an Economic Advisor at the UK Treasury and Development Associate at the Harvard Institute for International Development. 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. She is coauthor of *Strong Medicine: Creating Incentives for Pharmaceutical Research on Neglected Diseases*. 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.

Dan Levy

Dan Levy is Lecturer in Public Policy and Faculty Chair of the MPA Program at Harvard University's John F. Kennedy School of Government. He has served as a Senior Researcher at Mathematica Policy Research, where he has been involved in the evaluation of several social programs. His teaching focuses on quantitative methods and program evaluation. His research interests lie in the general area of social policy. He currently serves as the Deputy Project Director of the evaluation of the PATH program, a conditional cash transfer program in Jamaica. He has been involved in the evaluation of an after-school program, a methodological review of studies comparing the use of random assignment and quasi-experimental methods to estimate program impacts, and a technical assistance project with Mexico's Social Development Ministry (Sedesol). He has also served as a Research Affiliate of the Joint Center for Poverty Research, an Adjunct Faculty Member at Georgetown Public Policy Institute, and a consultant at the World Bank. He received his Ph.D. in Economics from Northwestern University and his B.A. from Universidad Metropolitana (Caracas, Venezuela). He is fluent in Spanish and French.

Leigh Linden

Leigh Linden is an Assistant Professor in both the Department of Economics and the School of International and Public Affairs at Columbia University. His research focuses on the ability of social services to improve the well being of children, especially in impoverished areas. He received his Ph.D. in Economics from the Massachusetts Institute of Technology in 2004 and his a B.S. in Mathematics and a B.A. in Economics in 1997 from the University of Texas at Austin.

Ben Olken

Benjamin Olken received his Ph.D. in Economics from Harvard in 2004, and is currently a Professor at Massachusetts Institute of Technology. His work focuses on empirical political economy questions in developing countries, with a particular emphasis on corruption. Most of his field work takes place in Indonesia, where he has conducted randomized field experiments and extensive data collection. He is an Affiliate of the Abdul Latif Jameel Poverty Action Lab, Faculty Research Fellow of the National Bureau of Economic Research, an Affiliate of the Bureau for Research and Economic Analysis of Development (BREAD), and a Research Affiliate of the Center for Economic Policy Research.

Marc Shotland

Marc is the Senior Research Manager and Director of Research at J-PAL. He holds an MPA/International Development degree from Harvard University's Kennedy School of Government and a Bachelors degree in Economics from Williams College. He first joined Professors Duflo and Banerjee in the summer of 2002 to run randomized evaluations of education interventions as a field research associate in India. In 2004 he joined the Poverty Action Lab's Cambridge office as a research manager. He left in 2006 to earn his Masters at Harvard before rejoining J-PAL in 2008 in his current position.

ABDUL LATIF JAMEEL
Poverty Action Lab



TRANSLATING RESEARCH INTO ACTION

Group One
Room 216

TA: Richard McDowell

Luis Alberro

Sub-Director
Mexican Ministry of Social Development
lalberro@gmail.com

Verna Jean Horgan

Executive Director
Hunt Institute for Engineering at SMU
vjhorgan@lyle.smu.edu

Stephanie Hunt

Co-Chair & Co-Founder
Hunt Institute for Engineering at SMU
steph-hunt@sbcglobal.net

Katherine Tait

Business Services Assitant
Boston Consulting Group
tait.katherine@bcg.com

ABDUL LATIF JAMEEL
Poverty Action Lab



TRANSLATING RESEARCH INTO ACTION

Group Two
Room 218

TA: Asha Stenquist

Esu Anahata

Co-Founder
The BARKA Foundation
inaandesu@barkafoundation.org

Marco Boggero

Head of Mission/Visiting Fellow
MSF Harvard
mboggero@fas.harvard.edu

Ines Kudo

Senior Projects Officer
The World Bank
ikudo@worldbank.org

Oluwasola Olanipekun

Senior Project Monitoring & Evaluation Officer
Affirmation of Rights of Persons With Disability
(ARPWD)
olanipekunsola@yahoo.com



Group Three
Room 220

TA: Emily Breza

Alawy Abdillahi

Monitoring and Evaluation Specialist
Women for Women International
aalawy@womenforwomen.org

Keith Amonlirdviman

Strategic Planning Manager
TechnoServe
keith@tns.org

Camila Alva

Research Assistant
International Food Policy Research Institute
camilachay@gmail.com

Celine Carbullido

Monitoring & Evaluation Officer
Women for Women International
ccarbullido@womeforwomen.org

Ashley Leblanc

Monitoring & Evaluation Officer
Women for Women International
aleblanc@womenforwomen.org

ABDUL LATIF JAMEEL
Poverty Action Lab



TRANSLATING RESEARCH INTO ACTION

Group Four
Room 222

TA: Eric Dodge

Jonathan Brooks

Resident Country Director
Millennium Challenge Corporation
brooksja@mcc.gov

Florenca Gabriele

Lecturer / Research Associate
Northeastern University
gabriele.f@husky.neu.edu

Mirian Lima

Head of Monitoring and Evaluation
Ministry of Finance, Cape Verde
Mirian.Lima@govcv.gov.cv

David Millet

Agro-Economist/Project Coordinator
Peasant Mouvement of Papaye
milletdavid@hotmail.com

Eleuthera Sa

Program Associate
Wellspring Advisors
esa@wellspringadvisors.com

ABDUL LATIF JAMEEL
Poverty Action Lab



TRANSLATING RESEARCH INTO ACTION

Group Five
Room 234

TA: Simone Schaner

Mary Ann Bates

Policy Associate
Jameel Poverty Action Lab
mary.ann.53@gmail.com

Radhika Joshi

Research Fellow
Lee Kuan Yew School of Public Policy
radhika.joshi@nus.edu.sg

Hammad Masood

Monitoring and Evaluation Specialist
UNICEF
hmasood@unicef.org

Makiko Omura

Associate Professor
Meijigakuin University
makiko@eco.meijigakuin.ac.jp

ABDUL LATIF JAMEEL
Poverty Action Lab



TRANSLATING RESEARCH INTO ACTION

Group Six
Room 236

TA: Reshma Hussam

Maria Aguirre

Researcher
Independent
maenith@hotmail.com

Bertha Briceno

Consultant
World Bank
bbriceno@worldbank.org

Diana Hincapie

PhD Student/Researcher
George Washington University
dianah@gwmail.gwu.edu

Dilip Rabha

Project Executive
Professional Assistance for Development Action
udayan_3046@yahoo.co.in

ABDUL LATIF JAMEEL
Poverty Action Lab



TRANSLATING RESEARCH INTO ACTION

Group Seven
Room 238

TA: Kamilla Gumedde

Varsha Harinath

Economist
Department of Trade & Industry, South Africa
VHarinath@thedti.gov.za

Neeta Misra

Post Doctorate Fellow
University of Cape Town
neeta.misra@uct.ac.za

Moleboheng Ntene

Economist
Department of Trade & Industry, South Africa
mntene@thedti.gov.za

Erin O'Brien

Researcher & Evaluator
Independent
erin.louise@gmail.com

Group Work Instructions

Groups are assigned to the follow locations

- E51-216, Group 1
- E51-218, Group 2
- E51-220, Group 3
- E51-222, Group 4
- E51-234, Group 5
- E51-236, Group 6
- E51-238, Group 7

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

The presentation should cover the following issues:

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

viii)

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

ix)

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?

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?

x)

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

xi)

How will the data be analyzed?

xii)

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



Case 1: 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 Raghavendra Chattopadhyay and Esther Duflo (2004a), *Econometrica* 72(5), 1409-1443.

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

Case 1: Women as Policymakers

India amended its federal constitution in 1992, devolving power to plan and implement development programs from the states to rural councils, or Gram Panchayats (Village Councils). The Village Councils 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 Village Council seats and Village Council chairperson positions for women. In most states, the schedule on which reserved seats and positions cycle among the Village Councils is determined randomly. This creates the opportunity to rigorously assess the impact of quotas on politics and government: Do the policies differ when there are more women in government? Do the policies chosen by women in power reflect the policy priorities of women? Since randomization was part of the Indian government program itself, the evaluation planning centered on collecting the data needed to measure impact. The researchers' questions were what data to collect, what data collection instruments to use, and what sample size to plan for.

Empowering the Panchayati Raj

Village Councils, known locally as 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 Village Councils 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 self-government, a system where every village would be “self-sustained and capable of managing its affairs.” Prime-minister Nehru advocated giving the Village Councils “great power,” so that rural Indians would “have a greater measure of real *swaraj* (*self-government*) in their own villages.”

Thus Article 40 of the constitution—adopted at independence—directs the states to ensure that the Village Councils “function as units of self-government.” Implementation guidelines recommended a three-tier system, with Village Councils as the grassroots unit.¹ Most states followed both the directive and the guidelines so that by the early 1950s they had formalized Village Councils. But in the 1960s, with no real power and no political and financial support from the federal government, the Village Councils 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 Village Council 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 chairperson has no veto power. But as the only councilor with a full-time appointment, the chairperson wields effective power.

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

Thinking about measurement and outcomes

The 73rd amendment aimed to decentralize the delivery of public goods and services essential for development in rural areas. The states were directed to delegate the power to plan and implement local development programs to the Village Councils. Funds still come from the central government but are no longer earmarked for specific uses. Instead, the Village Council decides which programs to implement and how much to invest in them. Village Councils can choose programs from 29 specified areas, including welfare services (for example, public assistance for 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 Village Councils are large and diverse. In West Bengal, for example, each has up to 12 villages and up to 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 Village Councils could ignore those groups' policy priorities. There were fears that the newly empowered Village Councils would undermine the development priorities of traditionally marginalized groups, such as women. To remedy this, the 73rd amendment included two mandates to ensure that investments reflected the needs of everyone in the Village Council.

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

The second mandate secures representation in the council for women. States are required to reserve at least a third of all council seats and Chairperson positions for women. Furthermore, the states have to ensure that the seats reserved for women are "allotted by rotation to different constituencies in a Panchayat [Village Council]" and that the chairperson positions reserved for women are "allotted by rotation to different Panchayats [Village Councils]." In other words, they have to ensure that reserved seats and chairperson positions rotate evenly within and among the Village Councils.

Randomized quotas in India: What can it teach us?

Your evaluation team has been entrusted with the responsibility to estimate the impact of quotas for women in the Village Councils. Your evaluation should address all dimensions in which quotas for women are changing local communities in India. What 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 quota policy. What needs did it address? What are the pros and cons of the policy? What can we learn from it?

2

The Abdul Latif Jameel Poverty Action Lab

@MIT, Cambridge, MA 02130, USA | @IFMR, Chennai 600 008, India | @PSE, Paris 75014, France

Case 1: Women as Policymakers

Discussion Topic 1: Gender quotas in the Village Councils

1. What were the main goals of the Village Councils?
2. Women are underrepresented in politics and government. Only 10 percent of India's national assembly members are women, compared to 17 percent worldwide.

Does it matter that women are underrepresented? Why and why not?

3. What were the framers of the 73rd amendment trying to achieve when they introduced quotas for women?

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

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

4. Debate the pros and cons of quotas. At the end of the debate, you should have a list of the pros and cons of quotas.
5. What evidence would you collect to strengthen the case of each party?

What data to collect

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

What are all the possible effects of quotas?

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

1. Brainstorm the possible effects of quotas: positive, negative, and no effects.
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 quotas will affect political participation of women, the indicator could be “number of women attending the General Assembly.”

Multiple outcomes are difficult to interpret, so define a hypothesis

Quotas for women could produce a large number of outcomes in different directions. For example, 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 quota policy, it will be very difficult to make sense of any result we find. Think of the following: if you take 500 villages and randomly assign them in your computer to a “treatment” group and a “control” group, and then run regressions to see whether the villages look different along 100 outcomes, would you expect to see some differences among them? Would it make sense to rationalize those results *ex-post*?

3

The Abdul Latif Jameel Poverty Action Lab

@MIT, Cambridge, MA 02130, USA | @IFMR, Chennai 600 008, India | @PSE, Paris 75014, France

Thinking about measurement and outcomes

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 quota 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 quotas are supposed to change policymaking, which will lead you to make predictions about which outcomes are affected.

Discussion Topic 2 continued...:

3. Suppose you had all the money and resources in the world and could collect data on every one of these indicators in communities with and without quotas, and compare them. How many indicators would you collect?
4. What might be some examples 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 continued...:

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

ABDUL LATIF JAMEEL

Poverty Action Lab



TRANSLATING RESEARCH INTO ACTION



Case 2: Learn to Read Evaluations

Evaluating the Read India Campaign

How to Read and Evaluate Evaluations

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

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

Why Learn to Read (L2R)?

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

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

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

Did the Learn to Read project work?

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

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

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

1

How to Read and Evaluate Evaluations

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

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

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

What comparison groups can we use? The following experts illustrate different methods of evaluating impact.

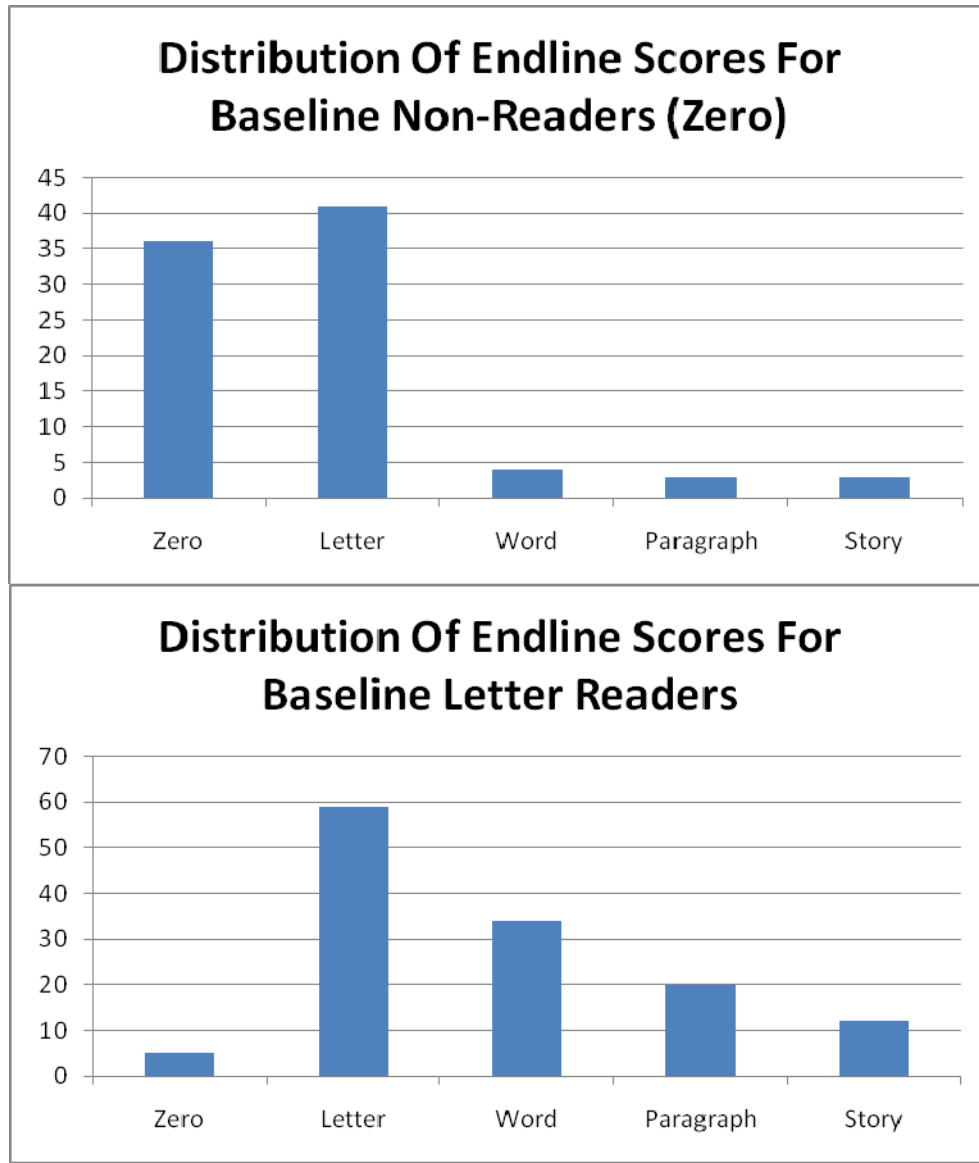
Estimating the impact of the Learn to Read project

Method 1:

News Release: Read India helps children Learn to Read.

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

Case 2: Learn to Read Evaluations



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

Discussion Topic 1:

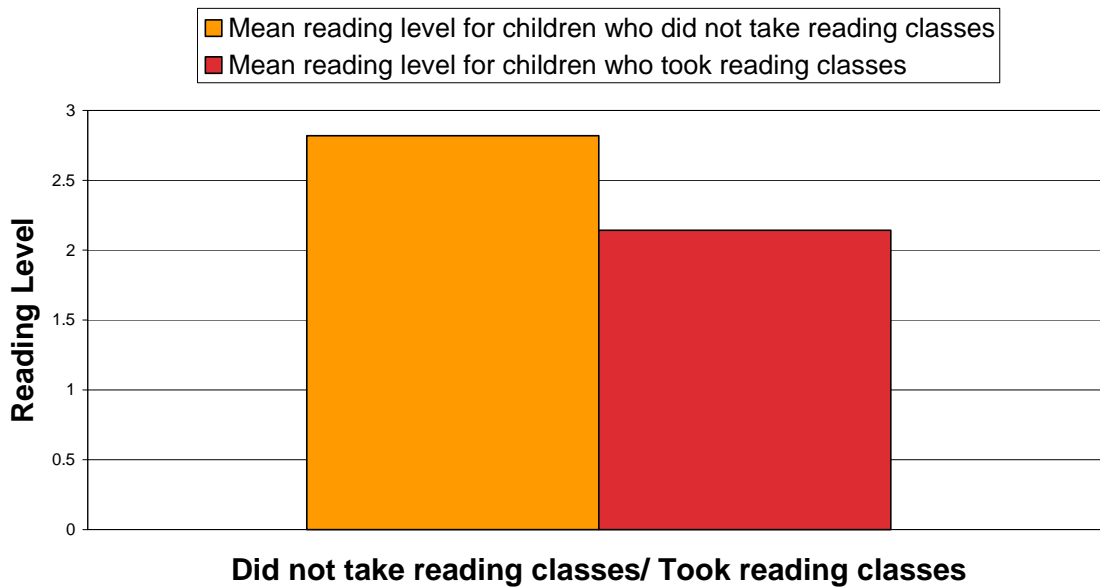
1. What type of evaluation does this news release imply?
2. What represents the counterfactual in this example?
3. What might be the problems with this type of evaluation (use concrete examples)?

Method 2:

Opinion: The “Read India” project not up to the mark

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

Comparison of reading levels of children who took reading classes Vs. reading levels of children who did not take them



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

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

Discussion Topic 2:

1. What type of evaluation is this opinion piece using?
2. What represents the counterfactual?
3. What might be the problem with this type of evaluation (use concrete examples)?

Method 3:

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

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

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

Discussion Topic 3:

1. What type of evaluation is this letter using?
2. What represents the counterfactual?
3. What might be the problem with this type of evaluation (use concrete examples)?

Method 4:

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

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

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

This report uses sophisticated statistical methods to measure the true impact of Pratham programs. We were concerned about other variables confounding previous results. We

How to Read and Evaluate Evaluations

therefore conducted a survey in these villages collecting information on child age, grade-level, and parents' education level and used those to predict child test scores.

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

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

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

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

Discussion Topic 4:

1. What type of evaluation is this report using?
2. What represents the counterfactual?
3. What might be the problem with this type of evaluation (use concrete examples)?

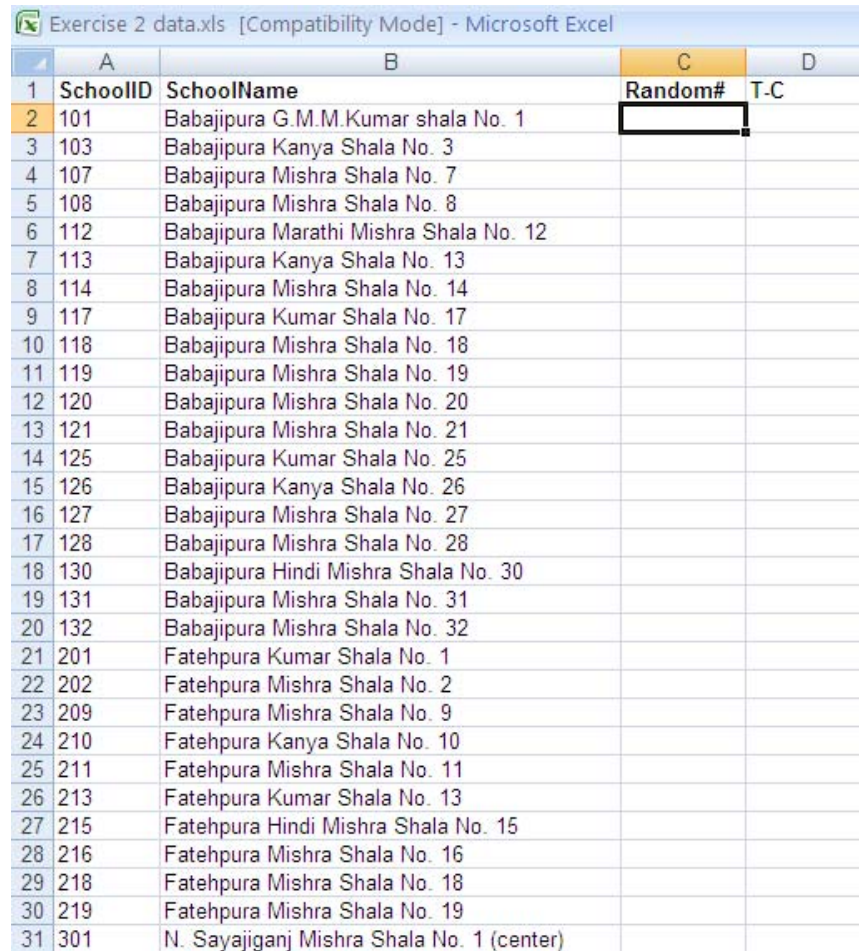
	Methodology	Description	Who is in the comparison group?	Required Assumptions	Required Data
Quasi-Experimental Methods	Pre-Post	Measure how program participants improved (or changed) over time.	Program participants themselves—before participating in the program.	The program was the only factor influencing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

Exercise 1: The mechanics of random assignment using MS Excel ®

Part 1: simple randomization

Like most spreadsheet programs MS Excel has a random number generator function. Say we had a list of schools and wanted to assign half to treatment and half to control

(1) We have all our list of schools.

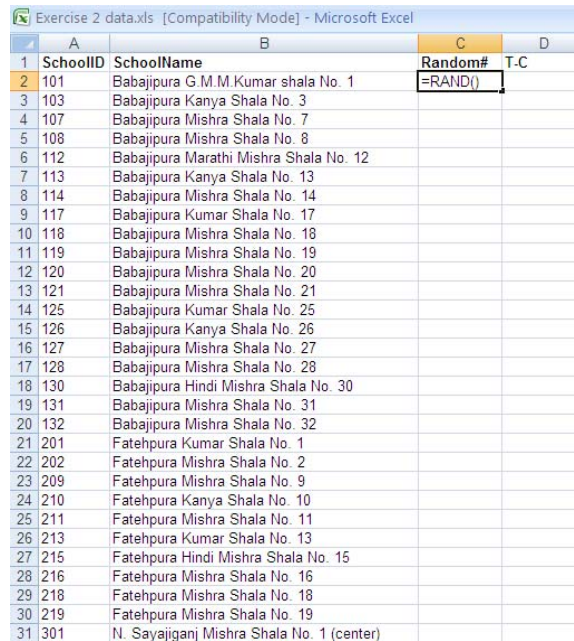


	A	B	C	D
1	SchoolID	SchoolName	Random#	T-C
2	101	Babajipura G.M.M.Kumar shala No. 1		
3	103	Babajipura Kanya Shala No. 3		
4	107	Babajipura Mishra Shala No. 7		
5	108	Babajipura Mishra Shala No. 8		
6	112	Babajipura Marathi Mishra Shala No. 12		
7	113	Babajipura Kanya Shala No. 13		
8	114	Babajipura Mishra Shala No. 14		
9	117	Babajipura Kumar Shala No. 17		
10	118	Babajipura Mishra Shala No. 18		
11	119	Babajipura Mishra Shala No. 19		
12	120	Babajipura Mishra Shala No. 20		
13	121	Babajipura Mishra Shala No. 21		
14	125	Babajipura Kumar Shala No. 25		
15	126	Babajipura Kanya Shala No. 26		
16	127	Babajipura Mishra Shala No. 27		
17	128	Babajipura Mishra Shala No. 28		
18	130	Babajipura Hindi Mishra Shala No. 30		
19	131	Babajipura Mishra Shala No. 31		
20	132	Babajipura Mishra Shala No. 32		
21	201	Fatehpura Kumar Shala No. 1		
22	202	Fatehpura Mishra Shala No. 2		
23	209	Fatehpura Mishra Shala No. 9		
24	210	Fatehpura Kanya Shala No. 10		
25	211	Fatehpura Mishra Shala No. 11		
26	213	Fatehpura Kumar Shala No. 13		
27	215	Fatehpura Hindi Mishra Shala No. 15		
28	216	Fatehpura Mishra Shala No. 16		
29	218	Fatehpura Mishra Shala No. 18		
30	219	Fatehpura Mishra Shala No. 19		
31	301	N. Sayajiganj Mishra Shala No. 1 (center)		

Incorporating Random Assignment into the Research Design

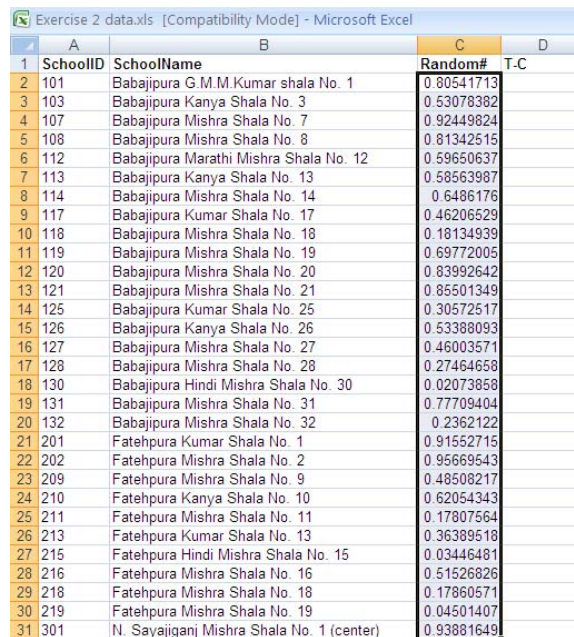
(2) Assign a random number to each school:

The function `RAND ()` is Excel's random number generator. To use it, in Column C, type in the following `= RAND()` in each cell adjacent to every name. Or you can type this function in the top row (row 2) and simply copy and paste to the entire column, or click and drag.



	A	B	C	D
1	SchoolID	SchoolName	Random#	T-C
2	101	Babajipura G.M.M.Kumar shala No. 1	=RAND()	
3	103	Babajipura Kanya Shala No. 3		
4	107	Babajipura Mishra Shala No. 7		
5	108	Babajipura Mishra Shala No. 8		
6	112	Babajipura Marathi Mishra Shala No. 12		
7	113	Babajipura Kanya Shala No. 13		
8	114	Babajipura Mishra Shala No. 14		
9	117	Babajipura Kumar Shala No. 17		
10	118	Babajipura Mishra Shala No. 18		
11	119	Babajipura Mishra Shala No. 19		
12	120	Babajipura Mishra Shala No. 20		
13	121	Babajipura Mishra Shala No. 21		
14	125	Babajipura Kumar Shala No. 25		
15	126	Babajipura Kanya Shala No. 26		
16	127	Babajipura Mishra Shala No. 27		
17	128	Babajipura Mishra Shala No. 28		
18	130	Babajipura Hindi Mishra Shala No. 30		
19	131	Babajipura Mishra Shala No. 31		
20	132	Babajipura Mishra Shala No. 32		
21	201	Fatehpura Kumar Shala No. 1		
22	202	Fatehpura Mishra Shala No. 2		
23	209	Fatehpura Mishra Shala No. 9		
24	210	Fatehpura Kanya Shala No. 10		
25	211	Fatehpura Mishra Shala No. 11		
26	213	Fatehpura Kumar Shala No. 13		
27	215	Fatehpura Hindi Mishra Shala No. 15		
28	216	Fatehpura Mishra Shala No. 16		
29	218	Fatehpura Mishra Shala No. 18		
30	219	Fatehpura Mishra Shala No. 19		
31	301	N. Sayajiganj Mishra Shala No. 1 (center)		

Typing `= RAND()` puts a 15-digit random number between 0 and 1 in the cell.



	A	B	C	D
1	SchoolID	SchoolName	Random#	T-C
2	101	Babajipura G.M.M.Kumar shala No. 1	0.80541713	
3	103	Babajipura Kanya Shala No. 3	0.53078382	
4	107	Babajipura Mishra Shala No. 7	0.92449824	
5	108	Babajipura Mishra Shala No. 8	0.81342515	
6	112	Babajipura Marathi Mishra Shala No. 12	0.59650637	
7	113	Babajipura Kanya Shala No. 13	0.58563987	
8	114	Babajipura Mishra Shala No. 14	0.6486176	
9	117	Babajipura Kumar Shala No. 17	0.46206529	
10	118	Babajipura Mishra Shala No. 18	0.18134939	
11	119	Babajipura Mishra Shala No. 19	0.69772005	
12	120	Babajipura Mishra Shala No. 20	0.83992642	
13	121	Babajipura Mishra Shala No. 21	0.85501349	
14	125	Babajipura Kumar Shala No. 25	0.30572517	
15	126	Babajipura Kanya Shala No. 26	0.53388093	
16	127	Babajipura Mishra Shala No. 27	0.46003571	
17	128	Babajipura Mishra Shala No. 28	0.27464658	
18	130	Babajipura Hindi Mishra Shala No. 30	0.02073858	
19	131	Babajipura Mishra Shala No. 31	0.77709404	
20	132	Babajipura Mishra Shala No. 32	0.2362122	
21	201	Fatehpura Kumar Shala No. 1	0.91552715	
22	202	Fatehpura Mishra Shala No. 2	0.95669543	
23	209	Fatehpura Mishra Shala No. 9	0.48508217	
24	210	Fatehpura Kanya Shala No. 10	0.62054343	
25	211	Fatehpura Mishra Shala No. 11	0.17807564	
26	213	Fatehpura Kumar Shala No. 13	0.36389518	
27	215	Fatehpura Hindi Mishra Shala No. 15	0.03446481	
28	216	Fatehpura Mishra Shala No. 16	0.51526826	
29	218	Fatehpura Mishra Shala No. 18	0.17860571	
30	219	Fatehpura Mishra Shala No. 19	0.04501407	
31	301	N. Sayajiganj Mishra Shala No. 1 (center)	0.93881649	

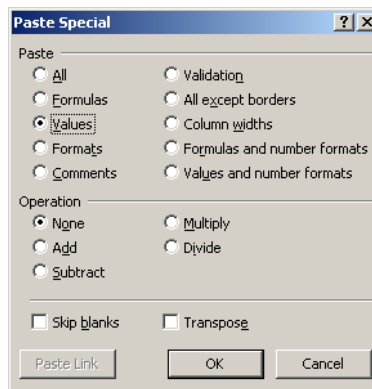
(3) Copy the cells in Column C, then paste the values over the same cells

The function, =RAND() will re-randomize each time you make any changes to any other part of the spreadsheet. Excel does this because it recalculates all values with any change to any cell. (You can also induce recalculation, and hence re-randomization, by pressing the key F9.)

This can be confusing, however. Once we've generated our column of random numbers, we do not need to re-randomize. We already have a clean column of random values. To stop excel from recalculating, you can replace the "functions" in this column with the "values".

To do this, highlight all values in Column C. Then right-click anywhere in the highlighted column, and choose Copy.

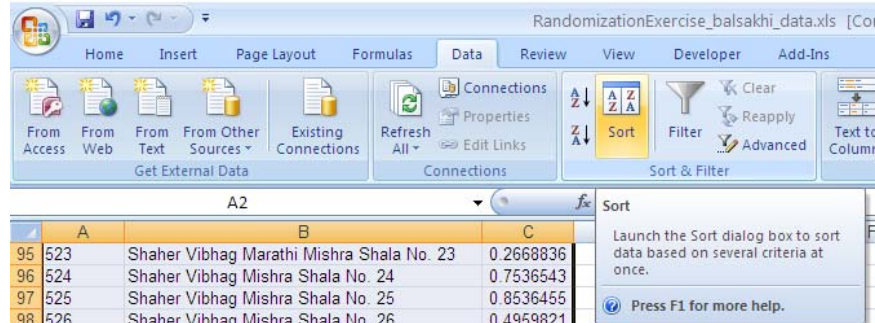
Then right click anywhere in that column and chose Paste Special. The "Paste Special" window will appear. Click on "Values".



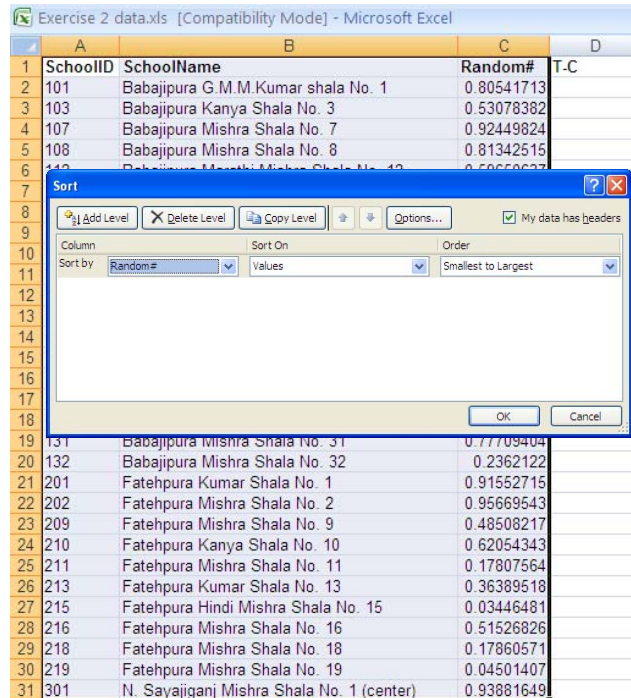
Incorporating Random Assignment into the Research Design

(4) Sort the columns in either descending or ascending order of column C:

Highlight columns A, B, and C. In the data tab, and press the Sort button:



A Sort box will pop up.



In the Sort by column, select “random #”. Click OK. Doing this sorts the list by the random number in ascending or descending order, whichever you chose.

Remedial Education: Evaluating the Balsakhi Program

There! You have a randomly sorted list.

	A	B	C	D
1	SchoolID	SchoolName	Random#	T-C
2	130	Babajipura Hindi Mishra Shala No. 30	0.02073858	
3	215	Fatehpura Hindi Mishra Shala No. 15	0.03446481	
4	219	Fatehpura Mishra Shala No. 19	0.04501407	
5	211	Fatehpura Mishra Shala No. 11	0.17807564	
6	218	Fatehpura Mishra Shala No. 18	0.17860571	
7	118	Babajipura Mishra Shala No. 18	0.18134939	
8	132	Babajipura Mishra Shala No. 32	0.2362122	
9	128	Babajipura Mishra Shala No. 28	0.27464658	
10	125	Babajipura Kumar Shala No. 25	0.30572517	
11	213	Fatehpura Kumar Shala No. 13	0.36389518	
12	127	Babajipura Mishra Shala No. 27	0.46003571	
13	117	Babajipura Kumar Shala No. 17	0.46206529	
14	209	Fatehpura Mishra Shala No. 9	0.48508217	
15	216	Fatehpura Mishra Shala No. 16	0.51526826	
16	103	Babajipura Kanya Shala No. 3	0.53078382	
17	126	Babajipura Kanya Shala No. 26	0.53388093	
18	113	Babajipura Kanya Shala No. 13	0.58563987	
19	112	Babajipura Marathi Mishra Shala No. 12	0.59650637	
20	210	Fatehpura Kanya Shala No. 10	0.62054343	
21	114	Babajipura Mishra Shala No. 14	0.6486176	
22	119	Babajipura Mishra Shala No. 19	0.69772005	
23	131	Babajipura Mishra Shala No. 31	0.77709404	
24	101	Babajipura G.M.M.Kumar shala No. 1	0.80541713	
25	108	Babajipura Mishra Shala No. 8	0.81342515	
26	120	Babajipura Mishra Shala No. 20	0.83992642	
27	121	Babajipura Mishra Shala No. 21	0.85501349	
28	201	Fatehpura Kumar Shala No. 1	0.91552715	
29	107	Babajipura Mishra Shala No. 7	0.92449824	
30	301	N. Sayajiganj Mishra Shala No. 1 (center)	0.93881649	
31	202	Fatehpura Mishra Shala No. 2	0.95669543	

(5) Sort the columns in either descending or ascending order of column C:

Because your list is randomly sorted, it is completely random whether schools are in the top half of the list, or the bottom half. Therefore, if you assign the top half to the treatment group and the bottom half to the control group, your schools have been “randomly assigned”.

In column D, type “T” for the first half of the rows (rows 2-61). For the second half of the rows (rows 62-123), type “C”

	A	B	C	D
1	SchoolID	SchoolName	Random#	T-C
2	130	Babajipura Hindi Mishra Shala No. 30	0.02073858	T
3	215	Fatehpura Hindi Mishra Shala No. 15	0.03446481	T
4	219	Fatehpura Mishra Shala No. 19	0.04501407	T
5	211	Fatehpura Mishra Shala No. 11	0.17807564	T
6	218	Fatehpura Mishra Shala No. 18	0.17860571	T
7	118	Babajipura Mishra Shala No. 18	0.18134939	T
8	132	Babajipura Mishra Shala No. 32	0.2362122	T
9	128	Babajipura Mishra Shala No. 28	0.27464658	T
10	125	Babajipura Kumar Shala No. 25	0.30572517	T
11	213	Fatehpura Kumar Shala No. 13	0.36389518	T
12	127	Babajipura Mishra Shala No. 27	0.46003571	T
13	117	Babajipura Kumar Shala No. 17	0.46206529	T
14	209	Fatehpura Mishra Shala No. 9	0.48508217	T
15	216	Fatehpura Mishra Shala No. 16	0.51526826	T
16	103	Babajipura Kanya Shala No. 3	0.53078382	T
17	126	Babajipura Kanya Shala No. 26	0.53388093	C
18	113	Babajipura Kanya Shala No. 13	0.58563987	C
19	112	Babajipura Marathi Mishra Shala No. 12	0.59650637	C
20	210	Fatehpura Kanya Shala No. 10	0.62054343	C
21	114	Babajipura Mishra Shala No. 14	0.6486176	C
22	119	Babajipura Mishra Shala No. 19	0.69772005	C
23	131	Babajipura Mishra Shala No. 31	0.77709404	C
24	101	Babajipura G.M.M.Kumar shala No. 1	0.80541713	C
25	108	Babajipura Mishra Shala No. 8	0.81342515	C
26	120	Babajipura Mishra Shala No. 20	0.83992642	C
27	121	Babajipura Mishra Shala No. 21	0.85501349	C
28	201	Fatehpura Kumar Shala No. 1	0.91552715	C
29	107	Babajipura Mishra Shala No. 7	0.92449824	C
30	301	N. Sayajiganj Mishra Shala No. 1 (center)	0.93881649	C
31	202	Fatehpura Mishra Shala No. 2	0.95669543	C

Incorporating Random Assignment into the Research Design

Re-sort your list back in order of school id. You'll see that your schools have been randomly assigned to treatment and control groups

	A	B	C	D
1	SchoolID	SchoolName	Random#	T-C
2	101	Babajipura G.M.M.Kumar shala No. 1	0.80541713	C
3	103	Babajipura Kanya Shala No. 3	0.53078382	T
4	107	Babajipura Mishra Shala No. 7	0.92449824	C
5	108	Babajipura Mishra Shala No. 8	0.81342515	C
6	112	Babajipura Marathi Mishra Shala No. 12	0.59650637	C
7	113	Babajipura Kanya Shala No. 13	0.58563987	C
8	114	Babajipura Mishra Shala No. 14	0.6486176	C
9	117	Babajipura Kumar Shala No. 17	0.46206529	T
10	118	Babajipura Mishra Shala No. 18	0.18134939	T
11	119	Babajipura Mishra Shala No. 19	0.69772005	C
12	120	Babajipura Mishra Shala No. 20	0.83992642	C
13	121	Babajipura Mishra Shala No. 21	0.85501349	C
14	125	Babajipura Kumar Shala No. 25	0.30572517	T
15	126	Babajipura Kanya Shala No. 26	0.53388093	C
16	127	Babajipura Mishra Shala No. 27	0.46003571	T
17	128	Babajipura Mishra Shala No. 28	0.27464658	T
18	130	Babajipura Hindi Mishra Shala No. 30	0.02073858	T
19	131	Babajipura Mishra Shala No. 31	0.77709404	C
20	132	Babajipura Mishra Shala No. 32	0.2362122	T
21	201	Fatehpura Kumar Shala No. 1	0.91552715	C
22	202	Fatehpura Mishra Shala No. 2	0.95669543	C
23	209	Fatehpura Mishra Shala No. 9	0.48508217	T
24	210	Fatehpura Kanya Shala No. 10	0.62054343	C
25	211	Fatehpura Mishra Shala No. 11	0.17807564	T
26	213	Fatehpura Kumar Shala No. 13	0.36389518	T
27	215	Fatehpura Hindi Mishra Shala No. 15	0.03446481	T
28	216	Fatehpura Mishra Shala No. 16	0.51526826	T
29	218	Fatehpura Mishra Shala No. 18	0.17860571	T
30	219	Fatehpura Mishra Shala No. 19	0.04501407	T
31	301	N. Sayajiganj Mishra Shala No. 1 (center)	0.93881649	C

Part 2: stratified randomization

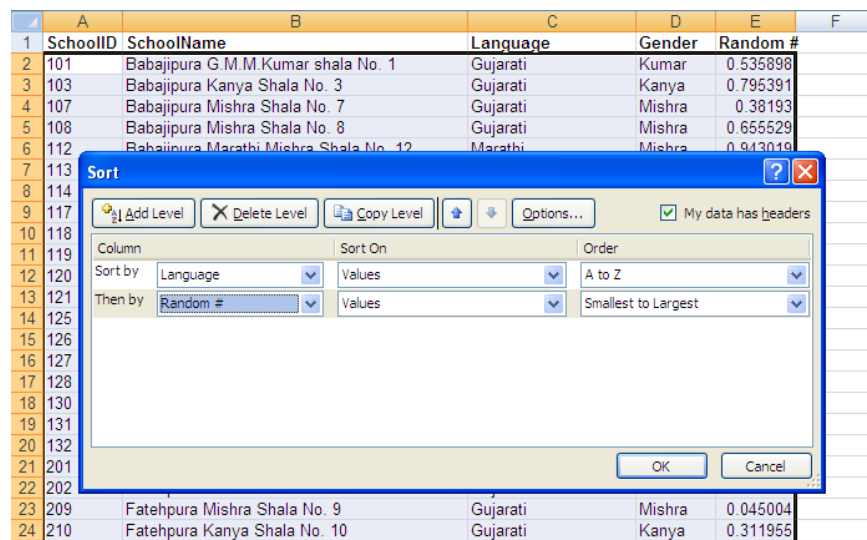
Stratification is the process of dividing a sample into groups, and then randomly assigning individuals within each group to the treatment and control. The reasons for doing this are rather technical. One reason for stratifying is that it ensures subgroups are balanced, making it easier to perform certain subgroup analyses. For example, if you want to test the effectiveness on a new education program separately for schools where children are taught in Hindi versus schools where children are taught in Gujarati, you can stratify by “language of instruction” and ensure that there are an equal number schools of each language type in the treatment and control groups.

(1) We have all our list of schools and potential “strata”.

Mechanically, the only difference in random sorting is that instead of simply sorting by the random number, you would first sort by language, and then the random number. Obviously, the first step is to ensure you have the variables by which you hope to stratify.

(2) Sort by strata and then by random number

Assuming you have all the variables you need: in the data tab, click “Sort”. The Sort window will pop up. Sort by “Language”. Press the button, “Add Level”. Then select, “Random #”.



(3) Assign Treatment – Control Status for each group.

Within each group of languages, type “T” for the first half of the rows, and “C” for the second half.

	A	B	C	D	E	F
100	132	Babajipura Mishra Shala No. 32	Gujarati	Mishra	0.8931975	C
101	615	Wadi Mishra Shala No. 15	Gujarati	Mishra	0.9142383	C
102	618	Wadi Kumar Shala No. 18	Gujarati	Kumar	0.9229356	C
103	408	Raopura Kanya Shala No. 8	Gujarati	Kanya	0.9285077	C
104	502	Shafer Vibhag Mishra Shala No. 2	Gujarati	Mishra	0.9549163	C
105	311	Sayajiganj Mishra Shala No. 11	Gujarati	Mishra	0.9595266	C
106	344	Sayajiganj Mishra Shala No. 44	Gujarati	Mishra	0.9688854	C
107	347	Sayajiganj Hindi Mishra Shala No. 47	Hindi	Mishra	0.0163449	T
108	332	Sayajiganj Hindi Mishra Shala No. 32	Hindi	Mishra	0.1528766	T
109	342	Sayajiganj Hindi Mishra Shala No. 42	Hindi	Mishra	0.2646791	T
110	215	Fatehpura Hindi Mishra Shala No. 15	Hindi	Mishra	0.3142377	T
111	326	Sayajiganj Hindi Mishra Shala No. 26	Hindi	Mishra	0.4291559	T
112	638	Wadi Hindi Mishra Shala No. 38	Hindi	Mishra	0.6772441	C
113	130	Babajipura Hindi Mishra Shala No. 30	Hindi	Mishra	0.7053783	C
114	315	Sayajiganj Hindi Mishra Shala No. 15	Hindi	Mishra	0.7955641	C
115	626	Wadi Hindi Mishra Shala No. 26	Hindi	Mishra	0.8918818	C
116	346	Sayajiganj Hindi Mishra Shala No. 46	Hindi	Mishra	0.9051467	C
117	303	N. Sayajiganj Marathi Mishra Shala No. 3	Marathi	Mishra	0.0354843	T
118	523	Shafer Vibhag Marathi Mishra Shala No. 23	Marathi	Mishra	0.1834626	T
119	409	Raopura Marathi Mishra Shala No. 9	Marathi	Mishra	0.7676874	T
120	611	Wadi Marathi Mishra Shala No. 11	Marathi	Mishra	0.8847497	T
121	329	Sayajiganj Marathi Mishra Shala No. 29	Marathi	Mishra	0.8992905	C
122	112	Babajipura Marathi Mishra Shala No. 12	Marathi	Mishra	0.9430188	C
123	327	Sayajiganj Marathi Mishra Shala No. 27	Marathi	Mishra	0.9515261	C
124	617	Wadi Marathi Mishra Shala No. 17	Marathi	Mishra	0.9648498	C

ABDUL LATIF JAMEEL

Poverty Action Lab



TRANSLATING RESEARCH INTO ACTION



Case 3: Extra Teacher Program

Designing an evaluation to answer
three key education policy questions

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

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

Case 3: Extra Teacher Program

Confronted with overcrowded schools and a shortage of teachers, in 2005 the NGO International Child Support Africa (ICS) offered to help the school system of Western Kenya by introducing contract teachers in 140 primary schools. Under its two year program, ICS provided funds to these schools to hire one extra teacher each year. In contrast to the civil servants hired by the Ministry of Education, contract teachers are hired locally by school committees. ICS expected this program to improve student learning by, among other things, decreasing class size and using teachers who are more directly accountable to the communities they serve. However, contract teachers tend to have less training and receive a lower monthly salary than their civil servant counterparts. So there was concern about whether these teachers were sufficiently motivated, given their compensation, or qualified given their credentials.

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

Over-crowded Schools

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

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

- 1) **Large class size:** Due to budget constraints, the rise in primary school enrollment has not been matched by proportionate increases in the number of teachers. (Teacher salaries already account for the largest component of educational spending.) The result has been very large class sizes, particularly in lower grades. In a sample of schools in Western Kenya, for example, the average first grade class in 2005 was 83 students. This is concerning because it is believed that small classes are most important for the youngest students, who are still acclimating to the school environment. The Kenyan National Union of Teachers estimates that the country needs an additional 60,000 primary school teachers in addition to the existing 175,000 in order to reach all primary students and decrease class sizes.
- 2) **Teacher absenteeism:** Further exacerbating the problem of pupil-teacher ratios, teacher absenteeism remains high, reaching nearly 20% in some areas of Kenya.

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

1

Designing an evaluation to answer three key policy questions

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

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

- 4) **Scarce school materials:** Because of the high costs of educational inputs and the rising number of students, educational resources other than the teacher are stretched, and in some cases up to four students must share one textbook. And an already over-burdened infrastructure deteriorates faster when forced to serve more children.
- 5) **Low completion rates:** As a result of these factors, completion rates are very low in Kenya with only 45.1% of boys and 43.3% of girls completing the first grade.

All in all, these issues pose new challenges to communities: how to ensure minimum quality of education given Kenya's budget constraints.

What are Contract Teachers?

Governments in several developing countries have responded to similar challenges by staffing unfilled teaching positions with locally-hired contract teachers who are not civil service employees. The four main characteristics of contract teachers are that they are: (1) appointed on annual renewable contracts, with no guarantee of renewed employment (unlike regular civil service teachers); (2) often less qualified than regular teachers and much less likely to have a formal teacher training certificate or degree; (3) paid lower salaries than those of regular teachers (typically less than a fifth of the salaries paid to regular teachers); and (4) more likely to be from the local area where the school is located.

Are Contract Teachers Effective?

The increasing use of contract teachers has been one of the most significant policy innovations in providing primary education in developing countries, but it has also been highly controversial. Supporters say that using contract teachers is an efficient way of expanding education access and quality to a large number of first-generation learners. Knowing that the school committee's decision of whether or not to rehire them the following year may hinge on performance, contract teachers are motivated to try harder than their tenured government counterparts. Contract teachers are also often more similar to their students in terms of geographic and cultural roots as well as socio-economic status. Opponents argue that using under-qualified and untrained teachers may staff classrooms, but will not produce learning outcomes. Furthermore the use of contract teachers de-professionalizes teaching, reduces the prestige of the entire profession, and

2

Case 3: Extra Teacher Program

reduces motivation of all teachers. Even if it helps in the short term, it may hurt efforts to recruit highly qualified teachers in the future.

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

The Extra Teacher Program Randomized Evaluation

In January 2005, International Child Support Africa initiated a two year program to examine the effect of contract teachers on education in Kenya. Under the program, ICS gave funds to 140 local school committees to hire one extra contract teacher to teach an additional first grade class. The purpose of this intervention was to address the first three challenges: class size, teacher accountability, and heterogeneity of ability. The evaluation was designed to measure the impact of class-size reductions, the relative effectiveness of contract teachers, and how tracking by ability would impact both low and high-achieving students.

Addressing Multiple Research Questions through Experimental Design

Different randomization strategies may be used to answer different questions. What randomization strategy could be used to evaluate the following questions? Concentrate on the appropriate unit (level) of randomization for each.

Discussion Topic 1: Testing the effectiveness of contract teachers

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

Discussion Topic 2: Looking at more general approaches of improving education

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

Discussion Topic 3: Addressing all questions with a single evaluation

4. Could a single evaluation explore all these issues at once?
5. What randomization strategy could do so?

Exercise 2: Understanding random sampling and the law of large numbers

In this exercise, we will visually explore random samples of different sizes from a given population. In particular, we will try to demonstrate that larger sample sizes tend to be more reflective of the underlying population.

- 1) Open the file “Exercise1_SamplingDistributions_NEW.xlsm”.
- 2) If prompted, select “Enable Macros”.
- 3) Navigate to the “Randomize” worksheet, which allows you to choose a random sample of size “Sample Size” from the data contained in the “control” worksheet.
- 4) Enter “10” for “Sample Size and click the “Randomize” button. Observe the distribution of the various characteristics between Treatment, Control and Expected. With a sample size this small, the percentage difference from the expected average is quite high for reading scores. Click “Randomize” multiple times and observe how the distribution changes.
- 5) Now, try “50” for the sample size. What happens to the distributions? Randomize a few times and observe the percentage difference for the reading scores.
- 6) Increase the sample size to “500”, “2000” and “10000”, and repeat the observations from step 5. What can we say about larger sample sizes? How do they affect our Treatment and Control samples? Should the percentage difference between Treatment, Control and Expected always go down as we increase sample size?

Sample size calculations

The Extra Teacher Program (ETP) case study discussed the concept of cluster randomized trials. The Balsakhi example used in the prior lecture introduced the concept of power calculations. In the latter, we were interested in measuring the effect of a treatment (balsakhis in classrooms) on outcomes measured at the individual level—child test scores. However, the randomization of balsakhis was done at the classroom level. It could be that our outcome of interest is correlated for students in the same classroom, for reasons that have nothing to do with the balsakhi. For example, all the students in a classroom will be affected by their original teacher, by whether their classroom is unusually dark, or if they have a chalkboard; these factors mean that when one student in the class does particularly well for this reason, all the students in that classroom probably also do better—which might have nothing to do with a balsakhi.

Therefore, if we sample 100 kids from 10 randomly selected schools, that sample is less representative of the population of schools in the city than if we selected 100 random kids from the whole population of schools, and therefore absorbs less variance. In effect, we have a smaller sample size than we think. This will lead to more noise in our sample, and hence larger standard error than in the usual case of independent sampling. When planning both the sample size and the best way to sample classrooms, we need to take this into account.

This exercise will help you understand how to do that. Should you sample every student in just a few schools? Should you sample a few students from many schools? How do you decide?

We will work through these questions by determining the sample size that allows us to detect a specific effect with at least 80% power. Remember power is the likelihood that when the treatment has an effect you will be able to distinguish it from zero in your sample.

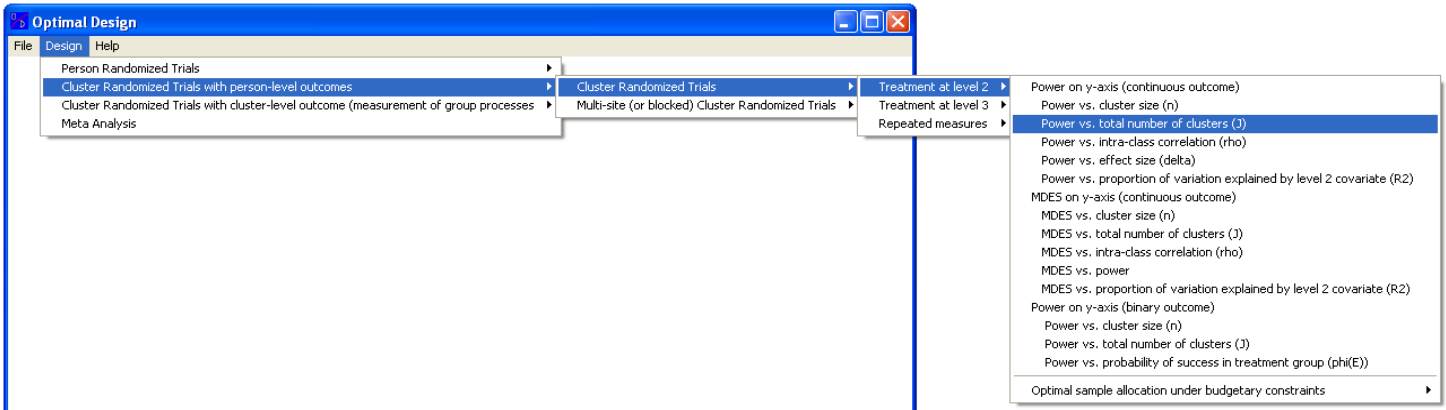
In this example, “clusters” refer to “clusters of children”—in other words, “classrooms” or “schools”. This exercise shows you how the power of your sample changes with the number of clusters, the size of the clusters, the size of the treatment effect and the Intraclass Correlation Coefficient. We will use a software program developed by Steve Raudebush with funding from the William T. Grant Foundation. You can find additional resources on clustered designs on their web site.

Section 1: Using the OD Software

First download the OD software from the website (a software manual is also available):

http://sitemaker.umich.edu/group-based/optimal_design_software

When you open it, you will see a screen which looks like the one below. Select the menu option “Design” to see the primary menu. Select the option “Cluster Randomized Trials with person-level outcomes,” “Cluster Randomized Trials,” and then “Treatment at level 2.” You’ll see several options to generate graphs; choose “Power vs. Total number of clusters (J).”



A new window will appear:



Select α (alpha). You'll see it is already set to 0.050 for a 95% significance level.

First let's assume we want to test only 40 students per school. How many schools do you need to go to in order to have a statistically significant answer?

Click on n , which represents the number of students per school. Since we are testing only 40 students per school, so fill in $n(1)$ with 40 and click OK.

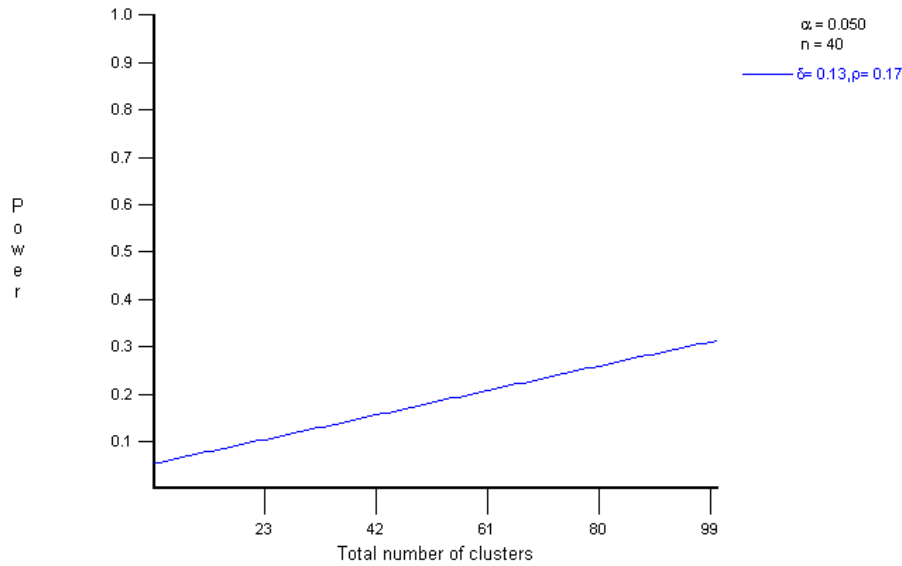
Now we have to determine δ (delta), the standard effect size (the effect size divided by the standard deviation of the variable of interest). Assume we are interested in detecting whether there is an increase of 10% in test scores. (Or more accurately, are uninterested in a detect less than 10%) Our baseline survey indicated that the average test score is 26, with a standard deviation of 20. We want to detect an effect size of 10% of 26, which is 2.6. We divide 2.6 by the standard deviation to get δ equal to 2.6/20, or 0.13.

Select δ from the menu. In the dialogue box that appears there is a prefilled value of 0.200 for delta(1). Change the value to 0.13, and change the value of delta (2) to empty. Select OK.

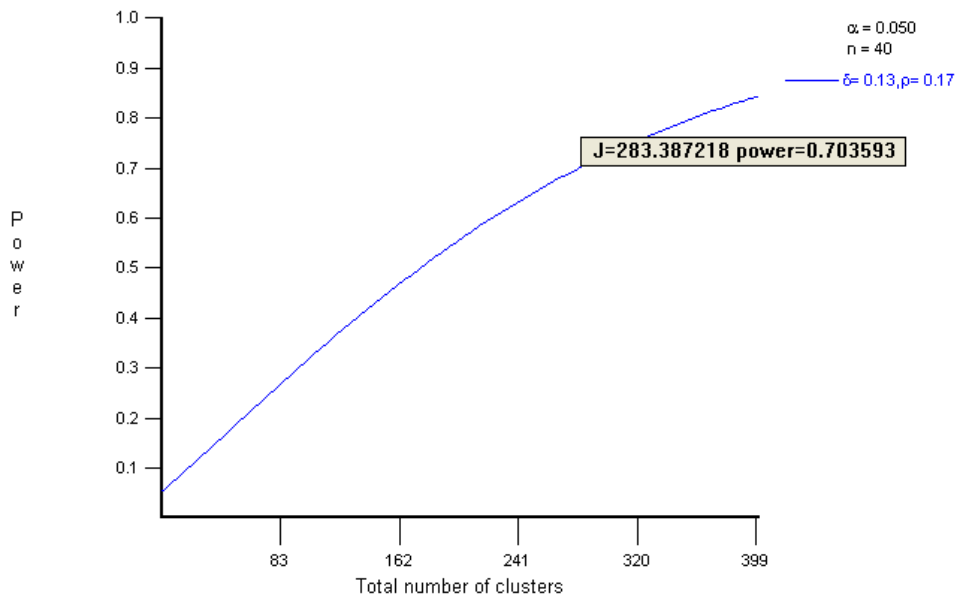
Finally we need to choose ρ (rho), which is the intra-cluster correlation. ρ tells us how strongly the outcomes are correlated for units within the same cluster. If students from the same school were clones (no variation) and all scored the same on the test, then ρ would equal 1. If, on the other hand, students from the same schools are in fact independent—and there was no differences between schools, then ρ will equal 0.

You have determined in your pilot study that ρ is 0.17. Fill in $\rho(1)$ to 0.17, and set $\rho(2)$ to be empty.

You should see a graph similar to the one below.



You'll notice that your x axis isn't long enough to allow you to see what number of clusters would give you 80% power. Click on the button to set your x axis maximum to 400. Then, you can click on the graph with your mouse to see the exact power and number of clusters for a particular point.



Exercise 3.1:
How many schools are needed to achieve 80% power? 90% power?

Now you have seen how many clusters you need for 80% power, sampling 40 students per school. Suppose instead that you only have the ability to go to 124 schools (this is the actual number that was sampled in the Balsakhi program).

Exercise 3.2:

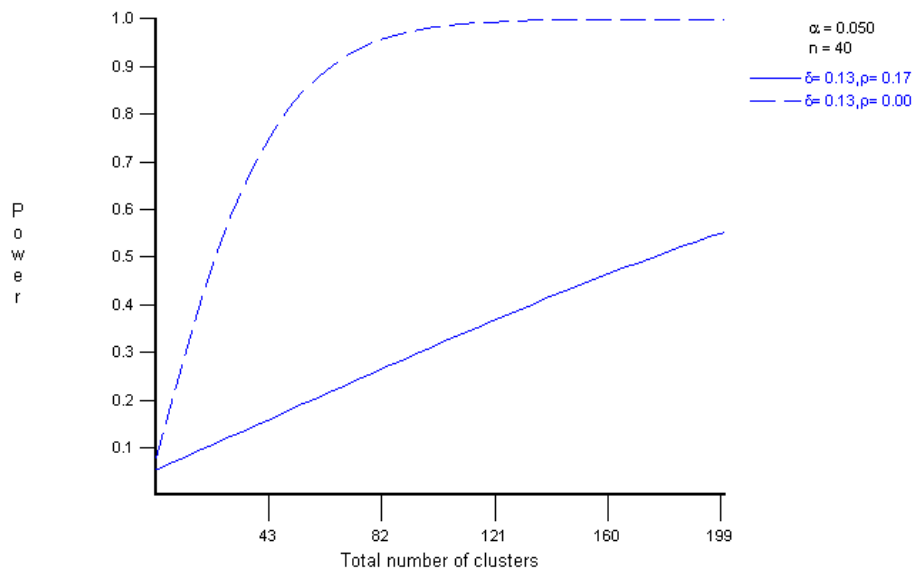
**How many children per school are needed to achieve 80% power? 90% power?
Choose different values for n to see how your graph changes.**


Finally, let's see how the Intraclass Correlation Coefficient (ρ) changes power of a given sample. Leave $\rho(1)$ to be 0.17 but for comparison change $\rho(2)$ to 0.0.

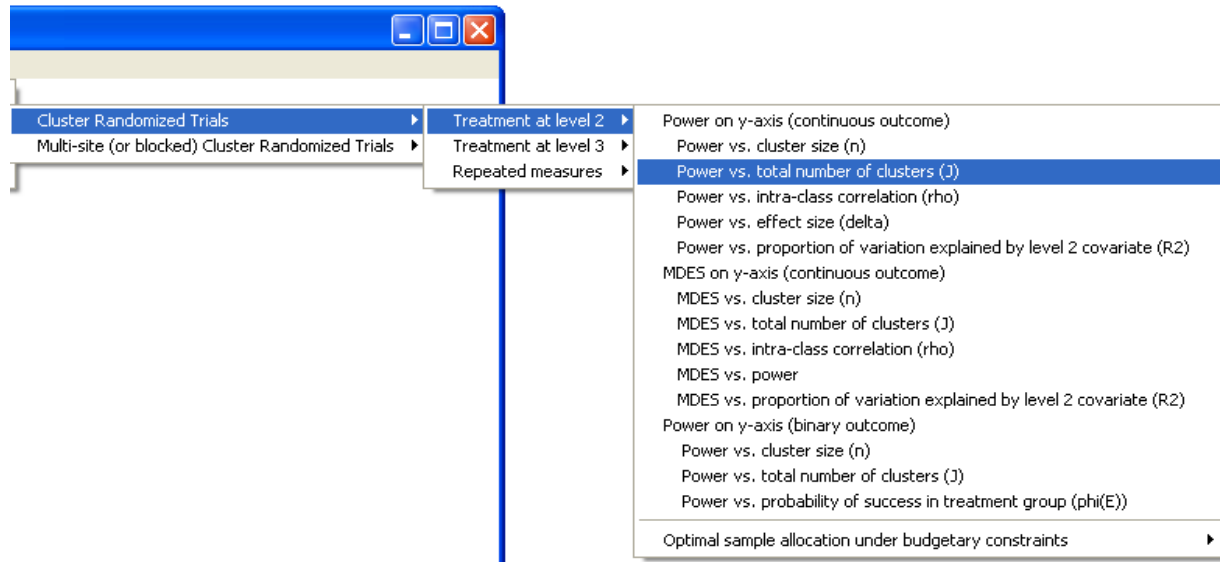
You should see a graph like the one below. The solid blue curve is the one with the parameters you've set - based on your pretesting estimates of the effect of reservations for women on drinking water. The blue dashed curve is there for comparison - to see how much power you would get from your sample if ρ were zero. Look carefully at the graph.

Exercise 3.3:

How does the power of the sample change with the Intraclass Correlation Coefficient (ρ)?



To take a look at some of the other menu options, close the graph by clicking on the  in the top right hand corner of the inner window. Select the Cluster Randomized Trial menu again.

**Exercise 3.4:**

Try generating graphs for how power changes with cluster size (n), intra-class correlation (ρ) and effect size (δ).

You will have to re-enter your pre-test parameters each time you open a new graph.

ABDUL LATIF JAMEEL

Poverty Action Lab



TRANSLATING RESEARCH INTO ACTION



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

Case 4: Deworming in Kenya

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.

1

The Abdul Latif Jameel Poverty Action Lab

@MIT, Cambridge, MA 02130, USA | @IFMR, Chennai 600 008, India | @PSE, Paris 75014, France

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 the 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?

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.

Worm Load	Pretest		Posttest	
	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

- What is the average pretest worm load for the treatment group?
 - What is the average pretest worm load for the comparison group?
 - Are the groups equivalent?
- What is the average posttest worm load for the treatment group?
 - What is the average posttest worm load for the comparison group?
 - What is the difference?
- Calculate the outcome differences at the beginning and at the end of the year.
 - Is this outcome difference an accurate estimate of the impact of the program?
 - If it is not accurate, does it overestimate or underestimate the impact?
- Because the treatment was treated, you expected there to be a difference between the groups at the end of the year.

 - If this difference is an effect, what is the source of attrition bias, if any?
 - How can you solve the problem to get a better estimate of program impact?
- 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)?
 - Calculate the impact of the program.
 - What is the size of the attrition bias?
- The PSPD also looked at school attendance rates and test scores.

 - Would differential attrition bias either of these outcomes?
 - Would the impact be underestimated or overestimated?
- In Case 1, you learned about other methods to estimate program impact, such as pre-post, simple difference, differences in differences, and multivariate regression.

 - Discuss if and how the issues explored above exist for each of these methods.
 - Are the threats to experimental integrity unique to randomization?

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 remained 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.

Worm Load	Pretest		Posttest	
	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 worms, so some parents 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 close 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.

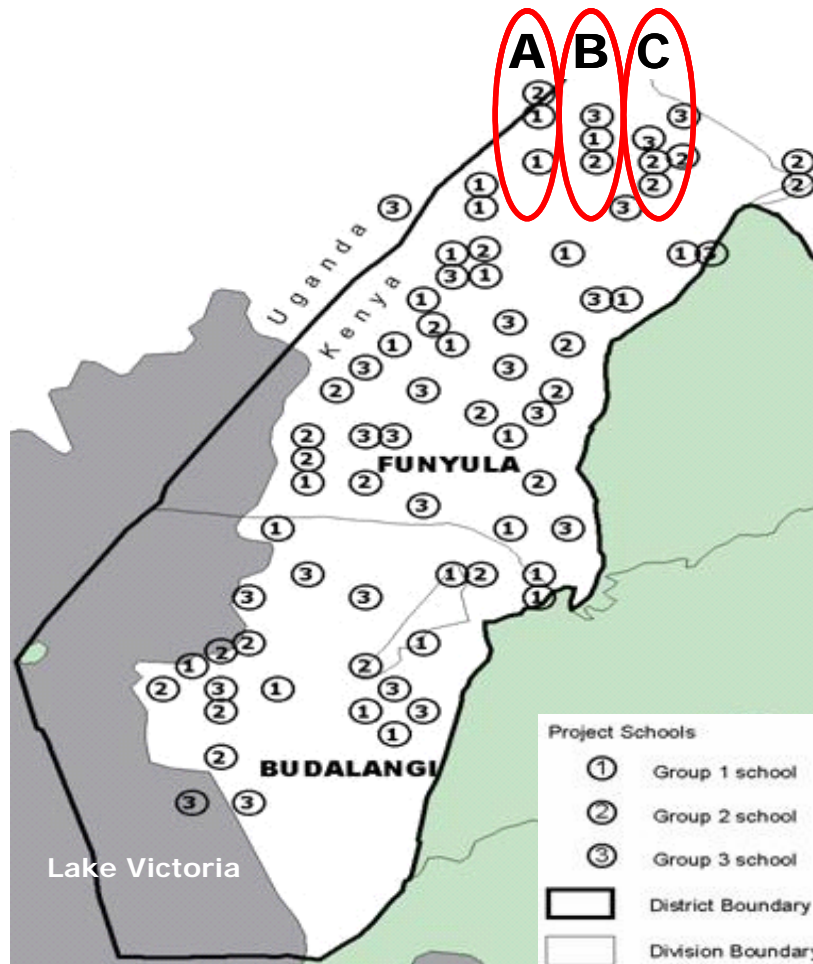
Each school has 100 children. Protocol compliance is 100 percent: all the children in treatment get treated and all the children in comparison do not get treated.

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

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.

This publication is in the public domain. Authorization to reproduce it in whole or in part for educational purposes is granted.

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/rigorousvid/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/RChapter4Final.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

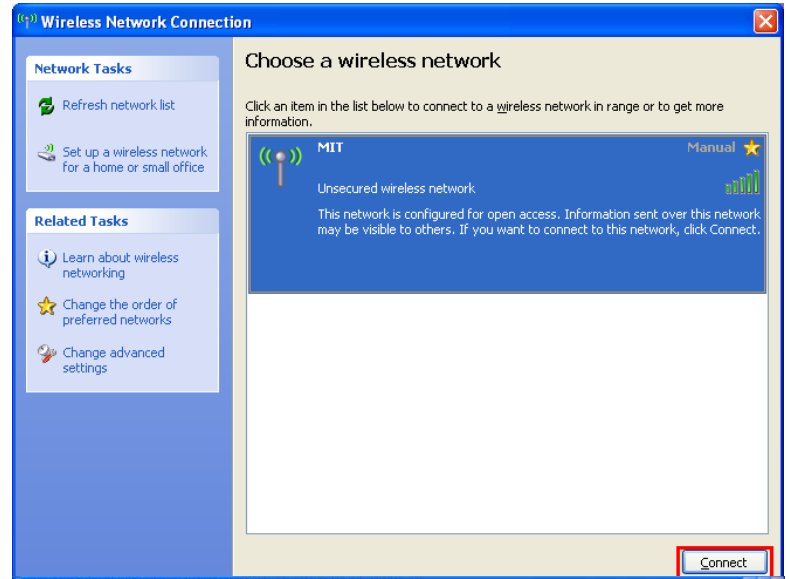
Connect (a school dropout prevention program for at-risk students), to evaluate the program's effects on dropping out of school – see <http://evidencebasedprograms.org/Default.aspx?tabid=163>; 56 schools containing nearly 6000 students were randomized in a trial of LifeSkills Training (a substance-abuse 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.

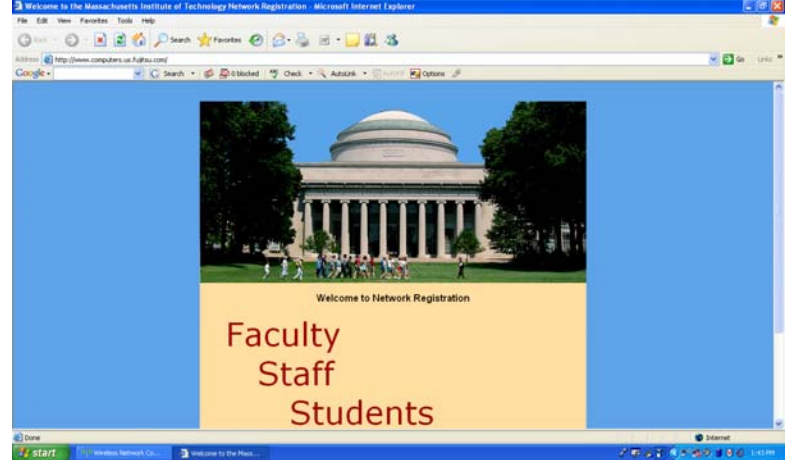
⁷ Similarly, a study of a crime prevention program that involves close police supervision of program participants should not use arrest rates as a measure of criminal outcomes, because the supervision itself may lead to more arrests for the intervention group.

MIT Wireless Instructions

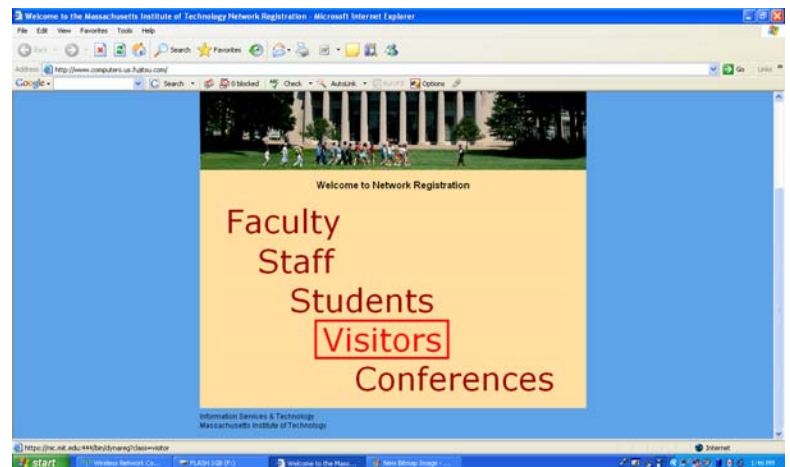
1) Search for available wireless networks. Select "MIT" and click connect. There are no passwords required here.



2) You will be automatically redirected to a screen that looks like this when you open a web browser.

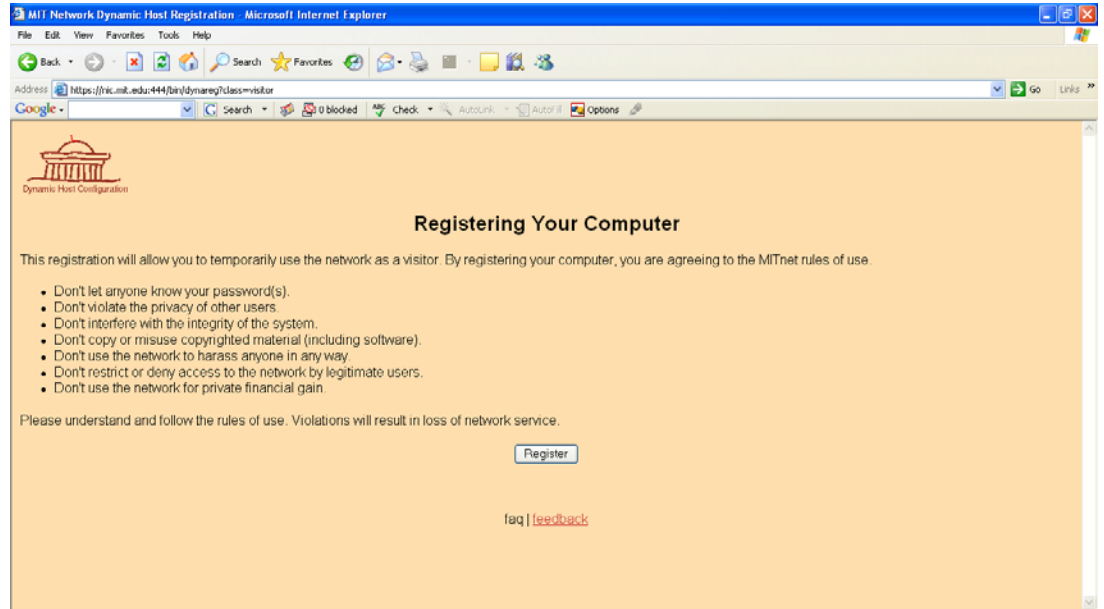


3) Select the visitor's option.

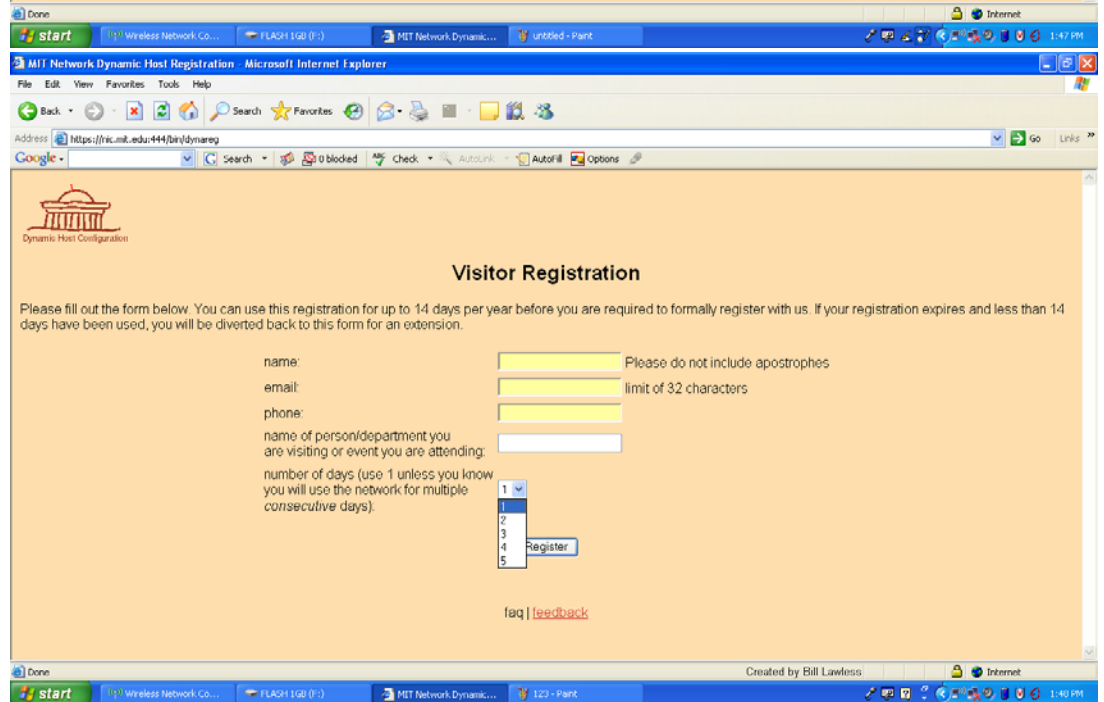


MIT Wireless Instructions

4) After reviewing the guidelines click the register button.



5. Fill out the form. Select the number of days that you will be here (5). Click the register button to submit.



6) Allow 15 minutes for information to replicate, and you should be all ready to surf the World Wide Web.