ABDUL LATIF JAMEEL
Poverty Action Lab
J-PAL EUROPE
TRANSLATING RESEARCH INTO ACTION

Evaluating Social Programs

Aarhus, September, 9-13, 2013

J-PAL Europe Executive Education Course



We are very grateful to the Trygfonden's Centre for Child Research of Aarhus University for its support and its help to organize this course.

TRYGFONDENS
BØRNEFORSKNINGSCENTER

AARHUS
UNIVERSITY

J-PAL EUROPE
Paris School of Economics - 66bis avenue Jean Moulin, 75014 Paris, France - T: +33 (0) 171 19 40 70
jpaleurope@povertyactionlab.org - http://www.povertyactionlab.org/europe

Table of Contents

1. Welcome Letter 3
2. Presentation of J-PAL 5
3. Course Schedule 7
4. Teachers presentation 9
5. Assistants presentation 11
6. Participants list 13
7. Group Presentation Guide 15
8. Work Groups 17
9. Case Study 1 – Getting Parents Involved 19
10. Case Study 2 – Get out the Vote 23
11. Case Study 3 – Counseling for Jobseekers 31
12. Case Study 4 – Deworming in Kenya 35
13. Exercise 1 – Understanding random sampling and the law of large numbers 43
14. Exercise 2 – The mechanics of random assignment using MS Excel 45
15. Exercise 3 – Power calculation 53
16. Bibliography 59
17. Checklist for Reviewing a RCT 61
18. Impact Evaluation Glossary 69
19. J-PAL and IPA Contacts 75
20. Notes 77

Welcome !

Each year, professors affiliated with J-PAL (the Jameel Poverty Action Lab) train dozens of people in the use of randomized evaluations. This year the Executive Education courses are also being held in Cape Town (South Africa), Cambridge (USA), Guatemala City (Guatemala), and New Delhi (India). J-PAL Europe organizes courses that annually alternate between English and French.

Due to the diversity of groups we welcome, each of these courses is unique.

Among you are people from Northern and Southern governments, international organizations, NGOs, and research institutes spanning 21 different countries around the world.

Some of you have already run research projects, others have a great deal of experience in operational work on the field. Some of you are statistics experts, while others are specialists on employment, education or agriculture areas. During this week, we will touch on all of these areas, enriching our course.

You have the opportunity to build relationships with other practitioners of randomized evaluations in different areas and contexts. We hope that common interests and new partnerships and projects will take shape during the week.

With this course, we hope to make you familiar with randomized evaluations: to understand what kind of valuable information they can provide, which constraints must be faced to implement them, when it is relevant to run one and when it is not. After the course, we remain at your disposal to bring you help and counseling, for questions about sample size or to help you to think about the evaluation of a specific program. Don't hesitate to contact us!

We wish you an interesting and productive week!



Hélène Giacobino
Directrice
J-PAL Europe

J-PAL: Researchers Against Poverty

Founded in 2003, the Abdul Latif Jameel Poverty Action Lab (J-PAL) is a network of researchers based at the Massachusetts Institute of Technology (MIT) in Cambridge, MA. In 2007, **J-PAL Europe** was officially launched at the Paris School of Economics, and **J-PAL South Asia** was founded at the Institute for Financial Management and Research in Chennai, India. In 2009, **J-PAL Latin America** was launched at the Pontifícia Universidad Católica in Santiago, Chile, in 2010 **J-PAL Africa** was founded at the University of Cape Town, in 2012 **J-PAL Southeast Asia** was launched at the University of Indonesia in Jakarta and in 2013 **J-PAL North America** was founded at the Massachusetts Institute of Technology (MIT).

J-PAL is a network of 83 researchers from all over the world. Since 2003 they have conducted more than 423 randomized evaluations of anti-poverty programs, of which 286 are already completed.

J-PAL works to improve the lives of the poor by ensuring that development policy is based on scientific evidence, collected through randomized evaluations.





What does J-PAL do?

J-PAL has three central objectives:

1. **Rigorous Evaluation of Development Projects:** J-PAL researchers are at the forefront of randomized evaluations, developing methodologies that allow an element of randomization to be introduced into programs in a way that is compatible with the constraints on the ground. As such, J-PAL researchers work on projects ranging from the evaluation of job search support programs in France, microfinance programs in rural Morocco, anti- corruption measures in India, and school feeding in Niger.

2. Capacity Building: Every year, J-PAL runs executive education courses in several locations around the world. These courses have trained hundreds of practitioners in more than 40 countries. Many of these practitioners have gone on to make significant contributions to randomized evaluations, either alone or in conjunction with J-PAL.

3. Diffusion of Results: J-PAL has also been very successful at promoting evidence-based policy through papers, conferences, seminars, and capacity building. Policymakers are increasingly using the evidence generated by J-PAL evaluations to guide their decisions. For example, findings on the impacts of school-based deworming have influenced government policy in Kenya and elsewhere.



J-PAL Funding

In 2005, J-PAL received a substantial gift from Mohammed Abdul Latif Jameel, an MIT alumnus and generous supporter of poverty alleviation initiatives around the world. The Poverty Action Lab was renamed in honor of his father, Abdul Latif Jameel.

Other donors and financial supporters include the Economic and Social Research Council, Swedish International Development Agency, DFID, The John D. and Catherine T. MacArthur Foundation, the Bill & Melinda Gates Foundation, the Hewlett Foundation, the National Institute of Health, the World Bank, Agence Française de Développement (AFD), the Haut Commissariat aux Solidarités Actives Contre la Pauvreté, the Institute Veolia Environnement, and the National Science Foundation.

Course Schedule

	Monday Sep, 9	Tuesday Sep, 10	Wednesday Sep, 11	Thursday Sep, 12	Friday Sep, 13
8:30 – 9:00	Registration/Breakfast				
9:00 – 10:30	Introduction to J-PAL: *Hélène GIACOBINO* Lecture 1: Evaluation: Why, What, When? *Michael Rosholm*	Case Study 2: Why Randomize?	Lecture 5: Sampling, Statistics, Sample size, Power *Roland Rathelot*	Lecture 6: Threats and Analysis *Elise Huillery*	Group Presentations
10:30 – 11:00	Coffee, Tea	Coffee, Tea		Coffee, Tea	Coffee, Tea
11:00 – 12:30	Case study 1: Theory of Change	Lecture 3: Why Randomize? *Marc Gurgand*	11:00 – 11:30: Coffee, Tea 11:30 – 12:30: Exercise 3	Group Project Work	Group Presentations Concluding Remarks
12:30 - 13:30	Lunch	Lunch	Lunch	Lunch	Lunch
13:30 – 15:00	Lecture 2: Measuring Impact: Outcomes & Indicators *Alessandro Tarozzi*	Lecture 4: How to Randomize? *Hélène Giacobino*	Case Study 4: Threats to analysis	Lecture 7: Start to Finish *Karen Macours*	
15:00 – 15:30	Coffee, Tea	Coffee, Tea	Coffee, Tea	Coffee, Tea	
15:30 – 18:00	Group Project Work	Case Study 3: How to Randomize? Exercises 1 & 2	Group Project Work	Finalize Group Work Presentation	
18:30 – 21:30		Dinner			

Teachers Presentation



Hélène GIACOBINO holds a Law degree and Bachelor degree in History and Film. After training as an Architect and then, as a Psychologist, she became a partner in a law firm, representing the firm in its international network for 15 years. She joined J-PAL Europe in 2009 as the Director of Strategy and Development and is now leading the office.
hgiacobino@povertyactionlab.org



Marc GURGAND is an Associate Professor at the Paris School of Economics. His research focuses on labor market policies, schooling and inequality in both developing and developed countries. He is currently conducting randomized evaluations of counseling schemes focused on the unemployed and welfare recipients. He also has a research program studying inequality in China. Marc Gurgand is a J-PAL Europe affiliate and its Scientific Director.
gurgand@pse.ens.fr



Élise HUILLERY is an Assistant Professor at the Department of Economics of Sciences Po. Her research focuses on policies addressing the lack of human capital (health, education, social capital) in developing countries and in France, with a special interest in understanding the psychological barriers to individual progression. She also has a research program on the colonial history and its long term impact in Africa.
elise.huillery@sciences-po.fr



Karen MACOURS is an Associate Professor at the Paris School of Economics and researcher at the Institut national de la recherche agronomique (INRA). Her current research focuses on conditional cash transfer programs, early childhood development, rural poverty, and agriculture.
karen.macours@parisschoolofeconomics.eu



Roland RATHELOT is a Researcher at the Centre de Recherche en Économie et Statistique (CREST). His areas of interest include labor economics, public economics and economics of immigration, with a particular focus on the spatial dimension. He is currently conducting randomized evaluations of counseling programs dedicated to the youth in France.
roland.rathelot@ensae.fr

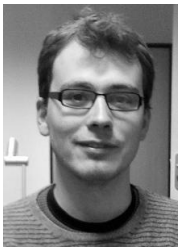


Alessandro TAROZZI is an Associate Professor at the Universitat Pompeu Fabra (UPF) and Barcelona Graduate School of Economics. His current research centers on factors that limit access and uptake of health-protecting technologies in developing countries. His work, which is mostly focused on India, also includes research on poverty estimation with missing data as well as on child nutritional status.
alessandro.tarozzi@upf.edu



Michael ROSHOLM is a Professor at the Aarhus University. He is a chairman of the Danish Economic Council and researches the effects of active labor market policies on individuals and firms, immigrants in labor market, health and employment.
rom@asb.dk

Assistants Presentation



Adrien BOUGUEN joined J-PAL Europe in October 2009. He is specialized in Economics of education. As a principal investigator he is currently working on several projects in various countries: in cooperation with the World Bank, he is part of an ongoing analyzes on the impact of a preschool construction program in Cambodia and with J-PAL Africa he is working on an impact evaluation about parental involvement.
abouguen@povertyactionlab.org



Axelle CHARPENTIER joined J-Pal Europe in October 2010. She is currently working on the randomized evaluation of a social mediation program to prevent school violence in France. She is also contributing to the impact evaluation of academic success programs (PREs) implemented as part of French urban social policy.
acharpentier@povertyactionlab.org



Clémence KIENY is currently working on the randomized evaluation of a project promoting the empowerment of young people in precarious situations in France. Before joining J-PAL, she worked as an intern at the International Trade Center in a program aiming at enhancing the transparency of global trade and market access rules.
ckieny@povertyactionlab.org



Bastien MICHEL joined J-PAL in 2008 and worked on numerous projects in France, Kenya and India, mostly in the Health and Education sectors. Here are a few projects he worked on: The Impact of VCT and Condom Distribution as HIV Prevention Strategies Among Youth in Kenya, Information and Referrals at the End of Middle School in France and Evaluating the Impact on Anemia of Making Double Fortified Salt Available in Bihar, India.
michel_bastien@hotmail.fr



Julie PERNAUDET Julie is currently working at CREST on a randomized evaluation of a health program for young people living in precarious situations in France. This evaluation aims at determining whether encouraging young people to meet a social worker and a doctor allows to increase their use of health care services. She just started a PhD on the role of parents' behavior on child cognitive and non-cognitive development at École Polytechnique and CREST.
julie.pernaudet@ensae.fr



Élise PÉSONEL joined J-PAL Europe in September 2011 and is currently working on two randomized evaluations of programs in France. She coordinates the evaluation of a program aiming to increase youth employment through a mentoring program for bursary master students. She is also working on an evaluation estimating the effects of a program facilitating French youth's access to apprenticeships and encouraging youth to complete them.

epesonel@povertyactionlab.org



Victor POULIQUEN joined J-PAL in 2008 and is currently working on three randomized evaluations focusing on education and health in Morocco, Ghana and Kenya. The first one looks at the impact of a conditional cash transfer program in education, the second at the effect of a scholarships program, and the third at the impact of different HIV/AIDS prevention programs.

vpouliquen@povertyactionlab.org

Participants List

Name	Title	Organization
ABADIA Laura	Policy Manager	J-PAL Europe, France
BAHRI Tarub	Fishery resources officer	Food and Agriculture Organization, Italy
BANK Lasse	Senior Advisor	The Danish National Labour Market Authority
BARINAGA Ester	Professor	Copenhagen Business School
BARO Alhassane Thierno	Magistrate	Court of Auditors, Senegal
BECH Stine	PhD Student	Aarhus University
BEKHOJ HANSEN Anders	Head of section	The Danish National Labour Market Authority
BEUCHERT Louise	PhD Student	Aarhus University
CHOOLHYE Hemasing	Analyst	Ministry of Finance and Economic Development, Mauritius
CHRISTENSEN Iben	Head of Office	The Danish National Labour Market Authority
DEY Subhasish	PhD Student	University of Manchester, United Kingdom
DYBDAL Line	Manager	Ramboll Management Consulting, Denmark
ERIKSEN Tine Louise Mundbjerg	PhD Student	Aarhus University
GEDEON ACHI Fiona	PhD Student	McGill University, Canada
GUSTAFSSON Line	Consultant	Ramboll Management Consulting, Denmark
HJORTSKOV LARSEN Morten	PhD Student	Aarhus University
HOGENKAMP Ineke	Senior Policy Advisor	Ministry of Foreign Affairs, Netherlands
HOYER ERIKSEN Lise	Special Advisor	The Danish National Labour Market Authority
KAMUGISHA Rick	Social Sientist	World Agroforestry Centre (ICRAF), Uganda
KARUHANGA Monica	Lecturer	Makerere University, Uganda
KRAEGPOTH Morten	PhD Student	Aarhus University
KRAMP Andre	Program Manager	Ministry of Education, Suriname

KUGONZA Jane	Dissemination Facilitator	World Agroforestry Centre (ICRAF), Uganda
LAMECH Simon	Head of Section	The Danish National Labour Market Authority
MASINO Serena	PhD Student	University of Manchester, United Kingdom
MENSINK Nico	Manager	FMO, Netherlands
MIZROKHI Elena	M&E Manager	MEDA, Morocco
MUBILA Maurice	Chief Statistician	African Development Bank, Tunisia
NAESBY Christian	Head of section	The Danish National Labour Market Authority
NJUGUNA Jane	Program Officer	Alliance for a Green Revolution in Africa (AGRA), Kenya
OOGARAH-BONOMAULLY Priyam	Senior Analyst	Ministry of Finance and Economic Development, Mauritius
OUELLET Rémi	Junior Technical Officer	Youth Employment Network (YEN), Switzerland
PAHALWANKHAN Faranaaz	Operations Officer	Ministry of Education, Suriname
PAYNE Lina	Evaluation Adviser	Department for International Development (DFID) UK
PEDERSEN Jonas Maibom	PhD Student	Aarhus University
PONTOPPIDAN Lise	Special Advisor	The Danish National Labour Market Authority
RODRIGO FLORES Rodimiro	Research Laboratory Manager	Fundacion Proacceso, Mexico
SCHROLL Michael	Planning and Evaluation Analyst	Office of Her Highness Sheika Moza Bint Nasser, Qatar
TIMSIT Juliette	Director	French American Charitable Trust, France
UNGUREANU Georgeta-Alina	Evaluation Officer	European Commission, Belgium
ZVALIONYTE Dovile	PhD Student	Vilnius University, Lithuania

Participants Email List:
At the end of the course, a contact list with your email will be sent to all of the participants. If you do not want to give your email address, please let us know and we will remove you from this list.

Help Desk for J-PAL Executive Education Course Alumni:
Last year, J-PAL launched the "RCT Help Online" (RHO). This moderated listserve is aimed at promoting an open discussion amongst participants. Participants will be automatically invited to the list by email after completion of the course.

Group Presentation

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:

1. Case studies and discussions
2. Preparation of group proposal

Case studies and Discussions



Each case study covers a specific set of topics corresponding to 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.

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 Friday, 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.

In order to help you for the preparation of the group proposal, a model in PowerPoint is available in your USB Key.



By Friday, you will output a 20-minute presentation (with an additional 10 minutes for questions and feedback).

The presentation should cover the following issues:

1. The objective and rationale of the evaluation—what is the question you are asking and why is it important or interesting?
2. Randomization design—how will the treatment and control groups be determined, and at what level will the randomization take place?
3. 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?
4. 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?
5. What are the risks to the integrity of the evaluation? How will you seek to minimize these?
6. How will the data be analyzed?
7. How will you use the results of the evaluation? How will the results impact future policy/programs?

Work Groups

Group 1

Assistant : Bastien Michel

Laura Abadia
Tarub Bahri
Rick Kamugisha
Monica Karuhanga
Jane Kugonza
Jane Njuguna

Group 2

Assistant : Axelle Charpentier

Fiona Gedeon Achi
Ineke Hogenkamp
Andre Kramp
Serena Masino
Faranaaz Pahalwankhan
Rodimiro Rodrigo Flores

Group 3

Assistant : Clémence Kieny

Stine Bech
Subhasish Dey
Elena Mizrokhi
Rémi Ouellet
Michael Schroll

Group 4

Assistant : Victor Pouliquen

Alhassane Thierno Baro
Hemasing Choolhye
Nico Mensink
Maurice Mubila
Priyam Oogarah-Bonomaully
Lina Payne

Group 5

Assistant : Adrien Bouguen

Lasse Bank

Anders Bekhoj Hansen

Tine Louise Mundbjerg Eriksen

Jonas Maibom Pedersen

Lise Pontoppidan

Georgeta-Alina Ungureanu

Group 6

Assistant : Julie Pernaudet

Iben Christensen

Line Dybdal

Line Gustafsson

Morten Kraegpoth

Christian Naesby

Dovile Zvalionyte

Group 7

Assistant : Élise Pésonel

Ester Barinaga

Louise Beuchert

Morten Hjortskov Larsen

Lise Hoyer Eriksen

Simon Lamech

Juliette Timsit

Case study 1

Getting Parents Involved



Program Theory, Measuring Outcomes

This case study is based on "Getting Parents Involved: A field Experiment in Deprived Schools"by Francesco Avvisati, Marc Gurgand, Nina Guyon, and Eric Maurin, CEPR Discussion Paper 8020, 2010.

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

Key Vocabulary

1. **Hypothesis:** a proposed explanation of the effects of a given intervention. Hypotheses should be made ex-ante, prior to the implementation of the intervention.
2. **Indicators:** metrics used to quantify and measure specific short-term and long-term effects of a program.
3. **Logical Framework:** a management tool used to facilitate the design, execution, and evaluation of an intervention. It involves identifying strategic elements (inputs, outputs, outcomes and impact) and their causal relationships, indicators, and the assumptions and risks that may influence success and failure.
4. **Theory of Change:** describes a strategy or blueprint for achieving a given long-term goal. It identifies the preconditions, pathways and interventions necessary for an initiative's success.

Due to problems of truancy and discipline, many children in industrialized societies graduate from school without mastering basic skills. The school district of Creteil (France) is a densely populated area with very poor socioeconomic indicators and high proportions of immigrants. In this setting, linguistic and social barriers along with financial and logistical constraints can prevent parents from paying closer attention to their children's education.

Increasing parental involvement has been widely touted as a means of overcoming difficulties in child learning and behavior. The program called "*La mallette des parents*" was designed to foster parental involvement through a series of monthly meetings with the school staff on how to successfully manage the transition from primary school to secondary school. These discussions provided parents of sixth graders (first year of middle school) with information on the French school system and guidelines on how to assist children with their homework.

Can parental involvement be used as a lever to improve educational outcomes in France? Does greater involvement of parents improve discipline and behavior? Do classroom interactions result in positive effects even for children whose parents don't attend the meetings?

The French Educational Environment

The French state-run educational system is highly centralized and schools have limited autonomy. All schools teach the same curriculum and employ teachers who are selected through national examinations. There is no tracking of students by ability and parents are not free to choose which school their children will attend.

Children enter middle school at age 11 or 12. For sixth graders, a typical week consists of 29 school hours, distributed across 9 different subjects, each taught by a different teacher. This is a major transition for pupils, after 5 years of primary school where each grade is taught by a single teacher.

The pool of students in the district of Creteil, where the program was implemented, is very heterogeneous both in abilities and in economic and cultural backgrounds. These eastern suburbs of Paris have large populations of first- and second-generation immigrants, many of which are

relatively poor (a recent survey showed that over 20 percent of the local population is composed of first-generation immigrants). These parents face many difficulties when trying to support their children throughout their school years: many speak little French, have limited understanding of the French educational system and work far away from their children's school. This lack of parental involvement might be the cause of problems such as truancy and indiscipline - especially in the poorest districts - and thus may contribute to many pupils not reaching the basic requirements of the curricula (OECD, 2010).

Informational Campaign for Parents

At the beginning of the academic year, all schools in our sample sent out informational leaflets to families of sixth graders offering them to register for a series of meetings organized by the school staff on how to successfully navigate the transition to middle school. Half of those schools were later randomly picked to implement the program.

The Scope of Discussions

The goal of these interactive meetings was to help parents understand the role of each member of the educational community and to help them develop positive involvement and attitudes towards their children's education. In order to lead these sessions, facilitators were given standard materials, including a DVD detailing the role of each staff member and documents explaining the functions of the various school offices. The first two sessions focused on how parents could help their children with homework, while the third session took place after the distribution of 1st term report cards, in order to help parents understand their child's results and to give them tips on how to go forward.

At the end of the third session, the principal asked participants whether they would like to participate in additional sessions on either parenting issues (in continuity with the first three meetings/debates) or on the use of internet-based tools to track their child's progress. Parents were also offered to attend French language sessions.

Your evaluation team has been entrusted with the responsibility of evaluating the campaign's impact on child learning and behavior. Your evaluation should address all dimensions in which informational campaigns for parents can affect cognitive and non-cognitive abilities of children. How might the meetings encourage greater involvement by parents? What are the most important outcomes to test? Which steps must occur in order for these changes to take place? What data should your team collect to evaluate the intervention?

Discussion Topic 1: Needs

1. Who is the target population?
2. What are the problems faced by these students?
3. Which characteristics of the French educational system make it particularly challenging for these students?
4. Which features of the home environment make it challenging?
5. What might be different in households of high-performing students?

Discussion Topic 2: Program Theory

1. What are the main characteristics (purpose, schedule, agenda, features...) of these informational meetings?
2. How might these meetings encourage parents to pay more attention to their children's education?
3. What are the potential challenges? Why might the program fail?

Discussion Topic 3: Outcomes and Indicators

1. What are the possible positive, negative and null effects of the intervention on child development and learning?
2. Please list all the indicators you would use to measure each of these potential outcomes.

Discussion Topic 4: Defining the Hypothesis

1. What might be some examples of key hypotheses you would test? Pick one.
2. Which indicators would you use to test your primary hypothesis?

Discussion Topic 5: Formalizing the Theory of Change

1. What are the steps or conditions that link the informational campaign for parents to the final outcomes?
2. Which indicators should you measure at each of these steps?
3. Distinguish the group of people on which you should collect data in order to measure the **outcome** of the intervention from that on which you should collect data in order to measure its **impact**.
4. Using the outcomes and conditions, draw a possible logical framework, linking the intervention to the final outcomes.

Case study 2

Get out the Vote



Do phone calls to encourage voting work?

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

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

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

Did the program work? How can we estimate its impact?

Voter turnout has been decaying since the 1960s

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

Vote 2002 get-out-the-vote Campaign

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

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

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

Did the Vote 2002 Campaign work?

What is required in order for us to measure whether a program worked, whether it had impact?

In general, to ask if a program works is to ask if the program achieves its goal of *changing certain outcomes* for its participants. To say, validly, that a program changes outcomes, we need to establish three things: (1) that outcomes have changed; (2) that the observed changes occurred among participants of the program and did not occur among non-participants; and (3) that it is not something else, some other event happening at the same time as the program, that drove the observed changes. In other words, we need to show that the program *causes* the observed changes.

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

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

The key challenge of program impact evaluation is constructing or mimicking the counterfactual. We typically do this by selecting a group of people that resemble the participants as much as possible but who did not participate in the program. This group is called the comparison group. Because we want to be able to say that it was the program and not some other factor that caused the changes in outcomes—condition (3) above—we want to be able to say that the only difference between the comparison group and the participants is that the comparison group did not participate in the program. We then estimate "impact" as the difference observed at the end of the program between the outcomes of the comparison group and the outcomes of the program participants.

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

That brings us back to our questions: Did the Vote 2002 Campaign work? What was its impact on voter turnout?

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

What comparison groups can we use?

Estimating the impact of the Vote 2002 Campaign

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

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

Method 1: Using a simple difference

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

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

	Voter turnout by group		*Impact Estimate*
	Reached	Not reached	
Method 1: Simple difference	64.5%	53.6%	10.8 pp*

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

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

Method 2: Using multivariate regression to control for inherent differences

Discussion Topic 2: Using multivariate regression

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

Characteristics of Reached and Not-Reached Groups

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

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

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

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

You could control for these differences by using a multivariate regression as follows: The participant and comparison groups are defined in the same way as in method 1. To estimate the impact of the program, you run a regression where the "dependent variable" is a zero/one variable indicating whether the person voted or not (i.e., 0 = did not vote, 1 = voted). The "key explanatory

variable" is a zero/one variable indicating whether the person received was reached (=1) or was not reached (=0). Potential differences in characteristics can be controlled for using other "explanatory variables" such as age, gender, newly registered voter, etc. The coefficient on the key explanatory variable (i.e., individual was reached) represents the "controlled" estimated impact of the program.

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

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

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

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

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

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

Discussion Topic 3: Using panel data

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

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

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

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

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

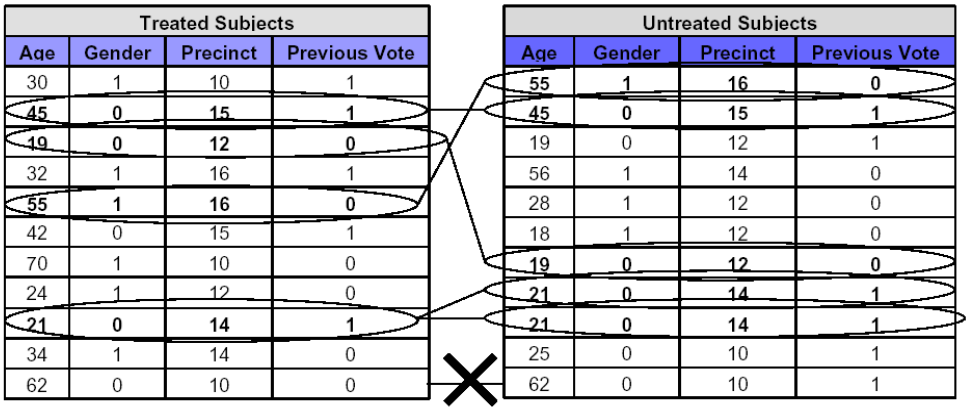
2. Given the information in the above table, would you expect that controlling for past voting behavior in method 2 would result in a higher or lower estimate of the impact of the Vote 2002 Campaign on voter turnout than the 6.1 pp found without controlling for it?

Method 4: Using matching

One way to estimate the impact of the Vote 2002 Campaign is to select as a comparison group a subset of non-participants who look similar to the participant group (people who were called and reached). To select this subset, researchers often employ a statistical procedure called *matching*. While there are many ways to do matching, it turns out that in this context it is possible to do *exact matching* for almost all the individuals in the sample. For each of the individuals reached, we can

select another individual who has the exact same characteristics (i.e., age, gender, etc.). In this way, the participant and comparison groups will have exactly the same observable characteristics. Figure 1 shows exact matching.

Figure 1: Exact Matching



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

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

Source: Arceneaux, Gerber, and Green (2004)

Discussion Topic 4: Exact Matching

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

Matching Analysis			
Number of Covariates matched on:	**Subset of Matched Reached**	**Subset of Matched Not-Reached Individuals**	***Impact***
4 (HH size, age, newly registered, state)	64.5%	60.8%	3.7 pp*
6 (HH size, age, newly registered, state in a competitive district, voted in 2000)	64.5%	61.5%	3.0 pp*
All *	65.9%	63.2%	2.7 pp*

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

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

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

Method 5: Using randomized experiments

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

Discussion Topic 5: Randomized Experiment

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

Compares the treatment and control groups on observable characteristics

	Treatment	Control	*Difference*
Voted in 2000	56.7%	56.4%	0.4 pp
Voted in 1998	22.7%	23.1%	-0.5 pp
Household Size	1.50	1.50	0.0 pp
Average age	52.0	52.2	-0.2 pp
% Female	54.6%	55.2%	-0.6 pp
% Newly registered	11.6%	11.7%	0.0 pp

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

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

Comparing voter turnout in the experimental treatment and the control groups

	Treatment (called)	Control (not called)	*Impact*
Simple Difference with randomization	58.2%	58.0%	0.2 pp
Difference after controlling for observable characteristics (multivariate regression)			0.2 pp

For the results to be comparable to the previous estimations, we need to adjust for the fact that not all individuals in the treatment group were reached. Indeed, since half of the treatment individuals were not reached, the simple difference result 0.2 pp is a diluted version of the impact of the program on those who were reached.

	Impact
Difference after adjusting for the fact that not all the people in the treatment group were reached ("Treatment Effect on the Treated" or TOT)*	0.4 pp

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

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

Comparing all five methods

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

Table 1: Comparing all five methods

Method	*Estimated impact*
Simple Difference	10.8 pp*
Multivariate Regression	6.1 pp*
Multivariate Regression with Panel Data	4.5 pp*
Matching (All Covariates)	2.8 pp*
Randomized experiment (treatment on the treated)	0.4 pp

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

(1) Result found when in addition to control for variables listed for the Multivariate Regressions method, we control for past voting behavior.

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

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

Discussion Topic 6: Selection bias

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

Case study 3

Counseling and Job Placement for Young Jobseekers



How to Randomize?

This case study is based on "Do Labor Market Policies have a Displacement Effect? Evidence from a Clustered Random Experiment." By Bruno Crepon , Esther Duflo, Marc Gurgand, Roland Rathelot, Philippe Zamora, Working Paper, 2011

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

Key Vocabulary

1. **Level of Randomization:** the level of observation (E.g. individual, household, school, village) at which treatment and control groups are randomly assigned.
2. **Spillovers:** individuals in the control group (or those not targeted for direct treatment) are indirectly affected by the treatment. In economics, these are called externalities. They can also be referred to as "contamination". Spillovers can be positive or negative.

Professional job counseling services are often discussed as a potential tool for helping educated young people find stable jobs. By connecting employers with job seekers, counseling agencies are thought to smooth the process of finding work and make better matches between employers and employees. Historically, the French government has taken it upon itself to provide these services. But how successful will this strategy be in solving France's problem of high unemployment—particularly among the youth? Even with these services, a sizable portion of those with college degrees have real difficulty finding a job. Some policymakers have suggested that *more intensive* forms of career counseling and support, in particular those provided by private agencies, could improve the efficiency of matching between employers and employees. Their proposals would reduce the role of the public sector in providing services for the unemployed, functionally handing over many of these core functions to the private sector.

If the government outsources this function to private employment agencies, will we see an improvement in job placement and job retention? What experimental designs could test the impact of this intervention?

The Problem of Chronic Unemployment

At the time this study, a large proportion of France's younger population was chronically unemployed, despite a generally healthy economy and the presence of public services to facilitate job placement. An estimated 25-32% of university graduates were unable to find stable work a full three years after graduation. While the government provided a safety net for many of the country's unemployed, such as money to cover basic necessities, to be eligible for such benefits a person must have been employed for at least 6 out of the 22 previous months, and must not have left the job out of their own free will. The job seekers selected for this study were generally in their mid-twenties, possessed vocational or university degrees, and had not had stable work for at least 6 months. Failing to meet the basic eligibility requirements, 69 percent of them were not receiving unemployment benefits. For them, the primary service had been counseling and placement services offered by the government.

Until 2005, the French Public agency ANPE (*Agence Nationale Pour l'Emploi*) was the sole provider of counseling and job placement for the unemployed French youth. The government compelled employers to communicate their vacant job announcements to ANPE, in order to make job placement swifter. However, the employment prospects of recent graduates remained dismal. In 2005, a law was passed that led to the proliferation of many private job placement firms. These private agencies were now allowed to openly propose their counseling and placement services

towards any jobseeker.

After the emergence of a private placement market, the government decided to increase the number of partnerships between the public operator and private actors. For this purpose, in 2007, the Ministry of Labor began delegating job placement for young graduates to more intensive counseling programs in private agencies in addition to their regular counseling program in the Public employment agency.

Details of the Program

Out of 30,000 unemployed youth identified in 10 regions of France, the government selected roughly 15,000 and assigned them to individual private agencies for counseling. The government did not prescribe a specific counseling structure, but it provided the agencies with incentives up to € 2100 per person counseled for meeting specific outcome targets:

- Payment 1: An eligible job seeker **enrolls** in their program.
- Payment 2: The job seeker **signs a job contract** at least 6 months in length.
- Payment 3: The job seeker **is still employed** 6 months after entering the job.

The government hand-selected the agencies that would be on its shortlist of service providers. Private counseling firms (for-profit and not-for-profit) were required to apply to the government to participate. (Outside of this program, these agencies served any individual jobseeker wishing to pay for their services). Upon winning the bid, they were committing themselves to serve all jobseekers assigned to them by the government under the three-payment incentive structure.

The agencies received the names of job seekers and contacted them to participate in two-stage counseling. The first stage focused on finding long-term employment (lasting at least six months). The second focused on stabilizing them in that new job.

The unemployed youth not selected to participate in the program still had the option of receiving counseling from the public employment agency, Pôle Emploi (or paying for the services of the private agencies on their own).

Addressing Key Experimental Issues through Evaluation Design

Different randomization strategies may be used to answer different questions. What strategies could be used to evaluate the following questions? How would you design the study?

Discussion Topic 1: Testing the effectiveness of private counseling

1. What is the relative effectiveness of private counseling versus regular government counseling? Who would be in the treatment and control groups, and how would they be randomly assigned to these groups?

Discussion Topic 2: Testing the effectiveness of for-profit and not-profit agencies

2. What is the relative effectiveness of for-profit private agencies versus non-profit private agencies? Who would be in the treatment and control groups, and how would they be randomly assigned to these groups?

Displacement Effects

Many economists argue that giving intensive job counseling to some individuals simply tips the scale in their favor, but does not increase job placements on net. In other words, it transfers job opportunities from individuals who do not receive counseling to those who do. Under this view, employment is a zero-sum game, and no counseling could increase employment.

In the context of an evaluation, the comparison (or control) group would be indirectly harmed by (their exemption from) the program, and would therefore no longer serve as a valid "counterfactual". This "negative spillover" could bias our estimate. If so, the experimental designs proposed above will be insufficient to measure the real effect of the program.

Discussion Topic 3: Managing Spillovers

1. How might spillovers undermine our analysis? In which direction could the bias be, and why?
2. What randomization strategy could you use to address this issue?

Discussion Topic 4: Measuring Spillovers

3. If you were interested in measuring whether spillovers exist, and specifically the impact of spillovers, how might you design the experiment differently?

Case study 4

Deworming in Kenya



Addressing 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

Key Vocabulary

1. **Phase-in Design**: a study design in which groups are individually phased into treatment over a period of time; groups which are scheduled to receive treatment later act as the comparison groups in earlier rounds.
2. **Equivalence**: groups are identical on all baseline characteristics, both observable and unobservable. Ensured by randomization.
3. **Attrition**: the process of individuals dropping out of either the treatment or comparison group over the course of the study.
4. **Attrition Bias**: statistical bias which occurs when individuals systematically drop out of either the treatment or the comparison group for reasons related to the treatment.
5. **Partial Compliance**: individuals do not "comply" with their assignment (to treatment or comparison). Also termed "diffusion" or "contamination."
6. **Intention to Treat**: the measured impact of a program comparing study (treatment versus control) groups, regardless of whether they actually received the treatment.
7. **Treatment on the Treated**: the measured impact of a program on participants who actually complied with treatment assignment.
8. **Externality**: an indirect cost or benefit incurred by individuals who did not directly receive the treatment. Also termed "spillover."

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 45,000 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 there can be external influences that can make them incomparable at the end of the program. Imagine we have a pile of seeds from 5 different plants. If we split this pile randomly into 2 bags, both bags should have the same composition of seeds. Suppose now that one of the bags gets perforated; the hole is small enough for only the smallest seed variety to pass through. What can we say about the composition of the two bags post this event? Are the two bags still comparable? Such events besides the program can happen between initial randomization and the end-line that 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, who typically have poorer sanitary habits, are particularly vulnerable: 400 million school-age children are chronically infected with intestinal worms.

Symptoms include listlessness, diarrhea, abdominal pain, and anemia. But worms affect more than the health of children. Heavy worm infections can impair children's physical and mental development, leading to poor attendance and performance in school.

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

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

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

Primary School Deworming Program

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

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

Evaluation design — the experiment as planned

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

schools in 1999, and Group 3 starting in 2001. Group 1 schools were the treatment group in 1998, while schools Group 2 and Group 3 were the comparison. In 1999 Group 1 and Group 2 schools were the treatment and Group 3 schools the comparison.

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

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

For the purpose of the following questions, we will look at results after the 1998 period.

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, both at the end of the program. 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?**

Managing attrition—when the groups do not remain equivalent

Attrition is when people 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 = score of 3
Medium worm infections = score of 2
Light infections = score of 1

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

Suppose protocol compliance is 100 percent: all children who are in the treatment get treated and none of the children in the comparison are treated. Children that were dewormed at the beginning of the school year (that is, children in the treatment group) end up with a worm load of 1 at the end of the year. The number of children in each worm-load category is shown for both the pretest and posttest.

	Pretest		*Posttest*	
Worm Load	Treatment	Comparison	Treatment	Comparison
3	5,000	10,000	0	10,000
2	5,000	10,000	0	10,000
1	5,000	10,000	15,000	10,000
Total children tested at school	15,000	30,000	15,000	30,000
Average				

1.
a. At pretest, what is the average worm load for each group?
b. At posttest, what is the average worm load for each group?
c. What is the impact of the program?
d. Do you need to know pretest values? Why or why not?

Suppose now that children who have a worm load of 3 only attend half the time and drop out of school if they are not treated. The number of children in each worm-load category is shown for both the pretest and posttest.

	Pretest		*Posttest*	
Worm Load	Treatment	Comparison	Treatment	Comparison
3	5,000	10,000	0	Dropped out
2	5,000	10,000	0	10,000
1	5,000	10,000	15,000	10,000
Total children tested at school	15,000	30,000	15,000	20,000
Average				

2.
a. At posttest, what is the new average worm load for the comparison group?
b. What is the impact of the program?
c. Is this outcome difference an accurate estimate of the impact of the program? Why or why not?
d. If it is not accurate, does it overestimate or underestimate the impact?
e. How can we get a better estimate of the program's impact?

3. Besides worm load, the PSDP also looked at outcome measures such as school attendance rates and test scores.
a. At posttest, what is the new average worm load for the comparison group?
b. Would the impacts on these final outcome measures be underestimated or overestimated?

4. In Case 2, you learned about other methods to estimate program impact, such as pre-post, simple difference, differences in differences, and multivariate regression.

a. Does the threat of attrition only present itself in randomized evaluations?

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 tutoring, 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. In any of these scenarios, 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 concern 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 deal with the complications that arise from partial compliance?

Discussion Topic 3: Managing partial compliance

Suppose all of the children from the poorest families have worm loads of 3. Their parents had also not paid the school fees. 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. While the children were allowed to stay in school during the year, because they had not paid school fees, these parents were reluctant to come to the school. Consequently, none of the children with worm loads of 3 were actually dewormed. Their worm load 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.

	Pretest		***Posttest***	
Worm Load	Treatment	Comparison	Treatment	Comparison
3	5,000	10,000	5,000	10,000
2	5,000	10,000	0	10,000
1	5,000	10,000	10,000	10,000
Total children tested at school	15,000	30,000	15,000	30,000

1. Calculate the impact estimate based on the original group assignments.

a. This is an unbiased measure of the effect of the program, but in what ways is it useful and in what ways is it not as useful?

You are interested in learning the effect of treatment on those actually treated ("treatment on the treated" (TOT) estimate).

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 and compare them to the comparison group.

a. Is this advice sound? Why or 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. Is this advice sound? Why or why not?

b. Would the impacts on these final outcome measures be underestimated or overestimated?

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

Exercise 1:

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 "Exercise A_SamplingDistributions.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?

Exercise 2:

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.

Exercise 2 data.xls [Compatibility Mode] - Microsoft Excel

	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)		

(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.

Exercise 2 data.xls [Compatibility Mode] - Microsoft Excel

	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.

Exercise 2 data.xls [Compatibility Mode] - Microsoft Excel

	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	

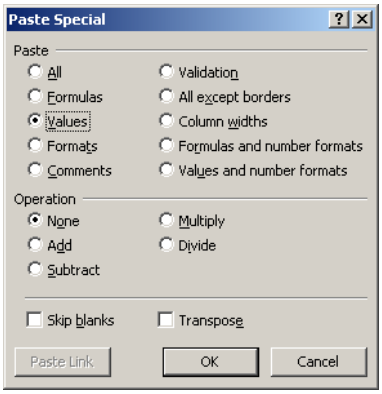
(3) Copy the cells in Colum C, then paste the values over the same cells

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".

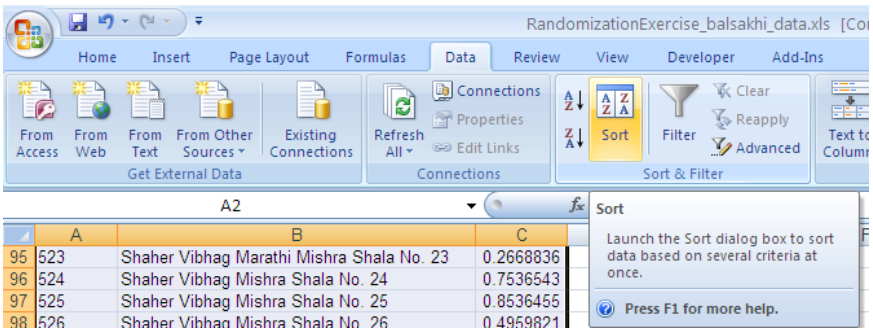
To do this, highlight all values in Column C. Then right-click anywhere in the highlighted column, and choose Copy.

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

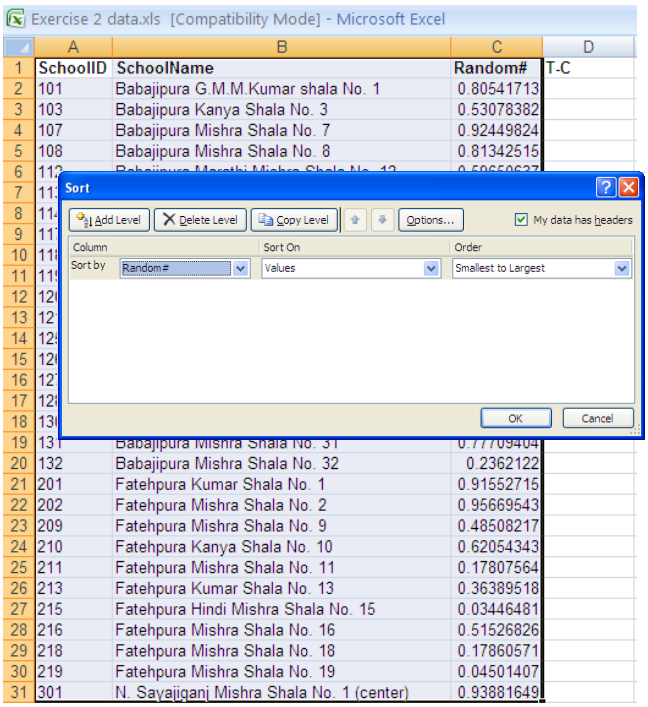


(4) Sort the columns in either descending or ascending order of column C:

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



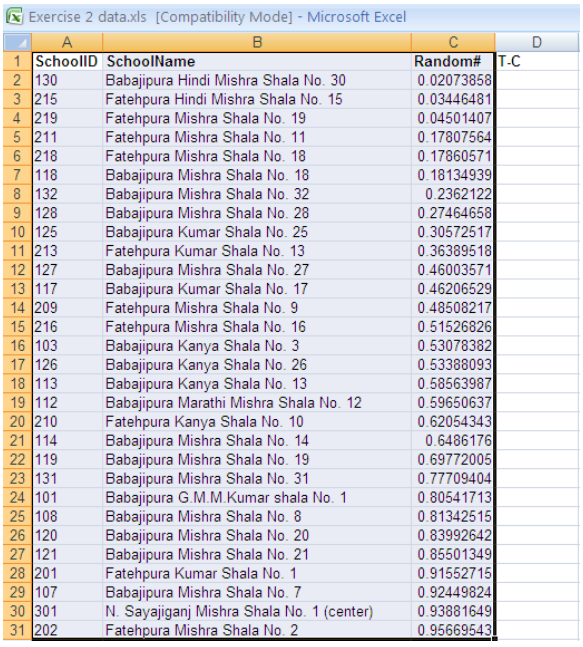
A Sort box will pop up.



	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	11			
8	11			
9	11			
10	11			
11	11			
12	12			
13	12			
14	12			
15	12			
16	12			
17	12			
18	13			
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	

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.

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"

Exercise 2 data.xls [Compatibility Mode] - Microsoft Excel

	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

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

Exercise 2 data.xls [Compatibility Mode] - Microsoft Excel

	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

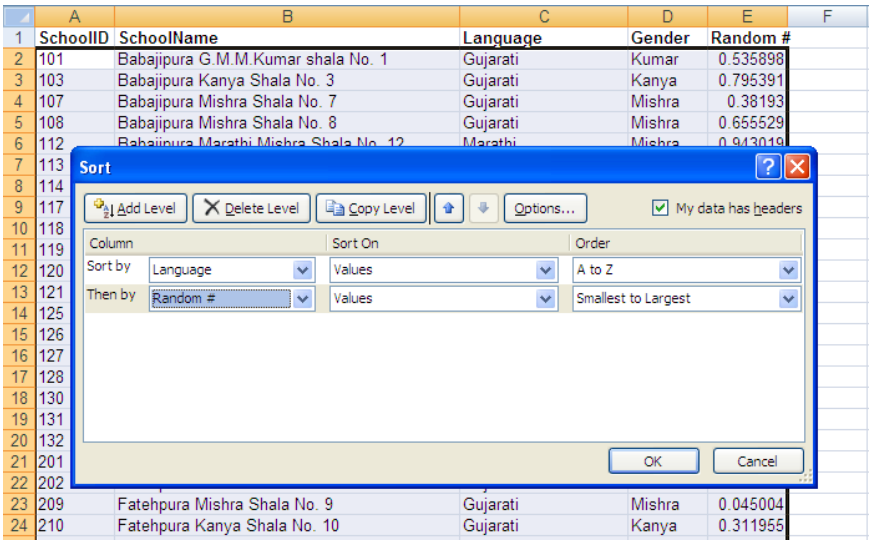
Stratification is the process of dividing a sample into groups, and then randomly assigning individuals within each group to the treatment and control. The reasons for doing this are rather technical. One reason for stratifying is that it ensures subgroups are balanced, making it easier to perform certain 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	Shaher 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	Shaher 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

Exercise 3:

Power Calculation Instructions

Key Vocabulary

1. **Power**: the likelihood that, when the program has an effect, one will be able to distinguish the effect from zero given the sample size.
2. **Significance**: the likelihood that the measured effect did not occur by chance. Statistical tests are performed to determine whether one group (e.g. the experimental group) is different from another group (e.g. comparison group) on the measurable outcome variables used in the evaluation.
3. **Standard Deviation**: a standardized measure of the variation of a sample population from its mean on a given characteristic/outcome. Mathematically, the square root of the variance.
4. **Standardized Effect Size**: a standardized measure of the [expected] magnitude of the effect of a program.
5. **Cluster**: the level of observation at which a sample size is measured. Generally, observations which are highly correlated with each other should be clustered and the sample size should be measured at this clustered level.
6. **Intra-cluster Correlation Coefficient**: a measure of the correlation between observations within a cluster; i.e. the level of correlation in drinking water source for individuals in a household.

Sample size calculations

The Extra Teacher Program will be used as an example to introduce the concept of power calculations and the concept of cluster randomized trials.

The 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 or Balsakhi in primary schools. Under its two year program, ICS provided funds to these schools to hire one extra teacher per school. In contrast to the civil servants hired by the Ministry of Education, contract teachers are hired locally by school committees. ICS expected this program to improve student learning by, among other things, decreasing class size and using teachers who are more directly accountable to the communities they serve. However, contract teachers tend to have less training and receive a lower monthly salary than their civil servant counterparts. So there was concern about whether these teachers were sufficiently motivated, given their compensation, or qualified given their credentials.

We were interested in measuring the effect of a treatment (balsakhis in classrooms) on outcomes measured at the individual level — child test scores.

In this experimentation, the randomization of balsakhis was done at the classroom level. However, 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.

Using the OD Software

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

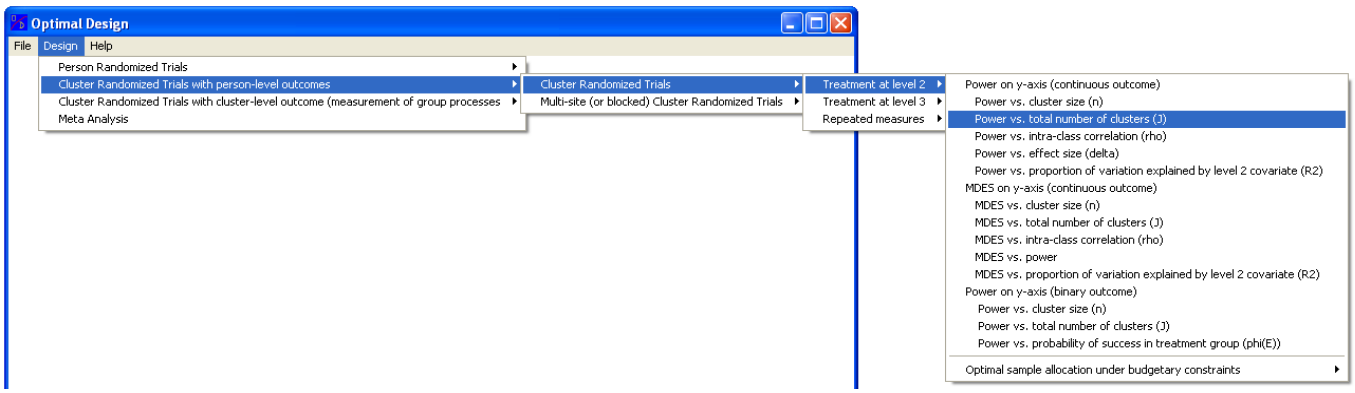
http://sitemaker.umich.edu/group-based/optimal_design_software

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,"then"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:



Now, change all the parameter used for sample size calculation:

- 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.

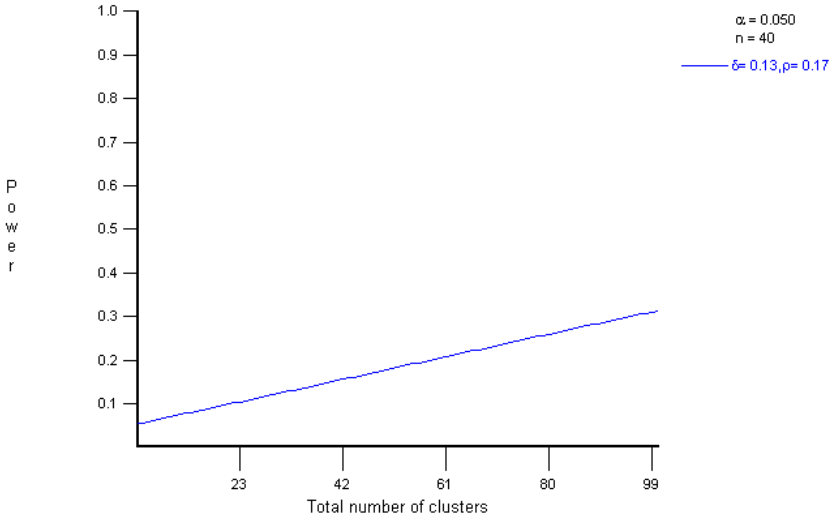
- Now we have to determine **δ** (delta), **the standard effect size** (the effect size divided by the standard deviation of the variable of interest). Assume we are interested in detecting whether there is an increase of 10% in test scores. (Or more accurately, are uninterested in a detect less than 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.

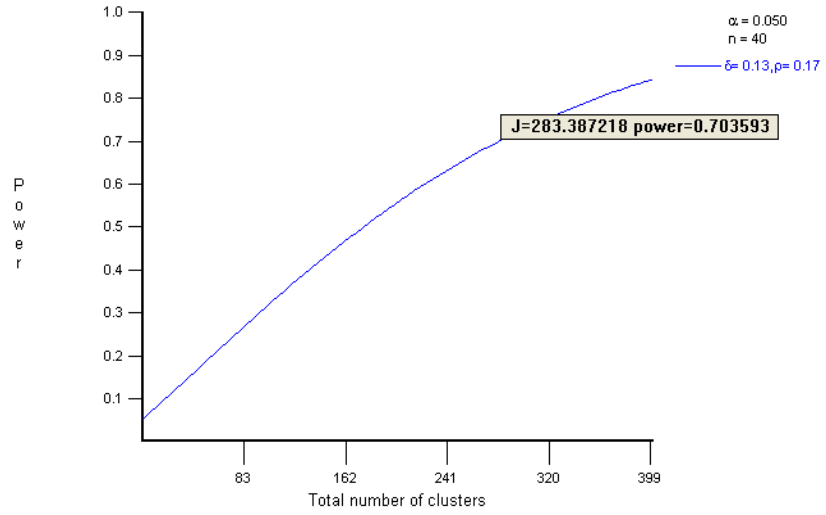
- Finally we need to choose ***ρ*** (rho), which is **the intra-cluster correlation**. ***ρ*** tells us how strongly the outcomes are correlated for units within the same cluster. If students from the same school were clones (no variation) and all scored the same on the test, then ***ρ*** would equal 1. If, on the other hand, students from the same schools are in fact independent—and there were 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 500. Then, you can click on the graph with your mouse to see the exact power and number of clusters for a particular point.



Exercise 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 300 schools.

Exercise 2:

How many children per school are needed to achieve 80% power?

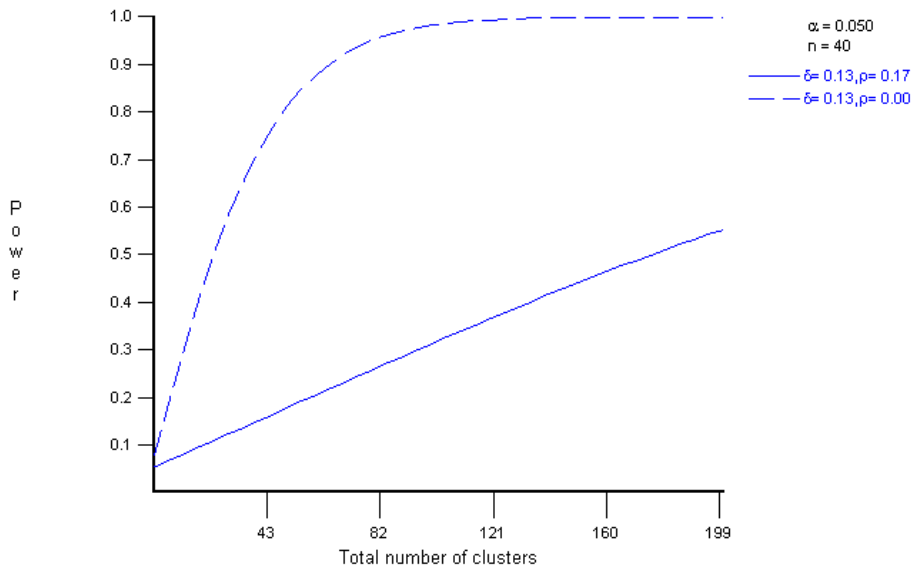
Choose different values for n to see how your graph changes.

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.

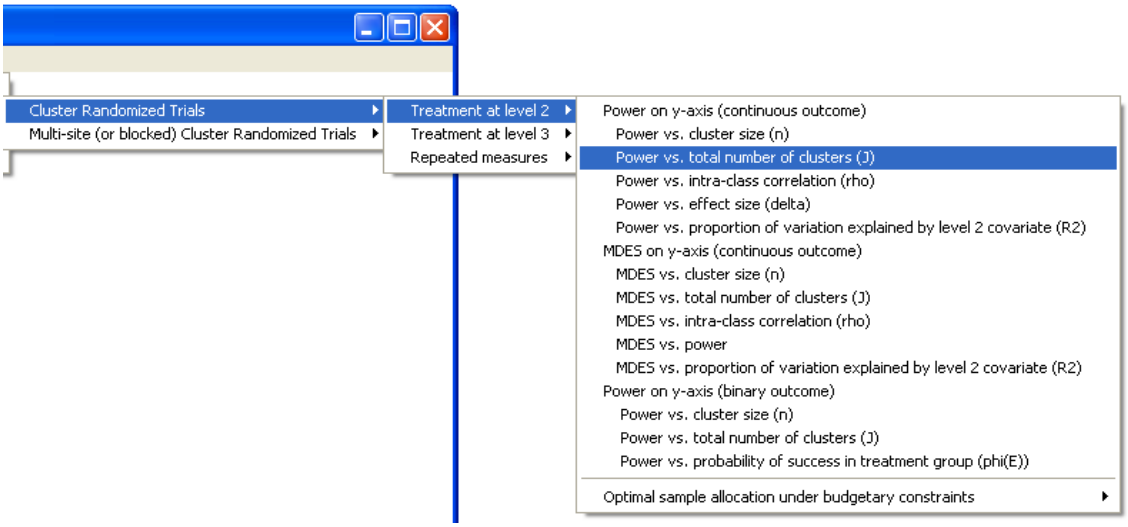
You should see a graph like the one below. The solid blue curve is the one with the parameters you've set - based on your pretesting estimates of the effect of **balsakhis in classrooms**. The blue dashed curve is there for comparison – to see how much power you would get from your sample if ***ρ***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

Exercise 3:

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 4:

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

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

Bibliography

To begin with...

Gertler, Martinez, Premand, Rawlings, Vermeersch : "Impact evaluation in practice", The World Bank, 2011: **http://tinyurl.com/kbgwv9p**

General Bibliography

Baird, Sarah, Joan Hamory, and Edward Miguel, 2008. "Tracking, Attrition, and Data Quality in the Kenyan Life Panel Survey Round 1" (KLPS-1), mimeo, George Washington University and UC Berkeley.

Banerjee, Abhijit, "Inside the Machine: Toward a New Development Economics," *Boston Review*, (March/April 2007).

Banerjee, Abhijit V. and Esther Duflo, "The Experimental Approach to Development Economics" Annual Review of Economics, (2009).

Banerjee, Abhijit V. and Esther Duflo, "Poor Economics", Public Affairs Books, (2011).

Deaton, Angus, "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development", NBER Working Paper, No. w14690, (January 2009).

Duflo, Esther "Field Experiments in Development Economics", Advances in Economic Theory and Econometrics, Eds. Richard Blundell, Whitney Newey, Torsten.Persson, *Cambridge University Press*, Volume 2(42), see also BREAD Policy Paper No. 002, 2005.

Duflo, Esther , Rachel Glennerster, Michael Kremer, "Using Randomization in Development Economics Research: A Toolkit" Handbook of Development Economics, Volume 4 (2008). *(a copy is in the USB key)*

Imbens, Guido "Better LATE than Nothing," NBER Working Paper No. w14896, (April 2009).

Imbens, Guido and Jeffrey M. Wooldridge, "Recent Developments in the Econometrics of Program Evaluation", *Journal of Economic Literature*, Vol. 47, No. 1, (March 2009), pp. 5-86.

Kling, Jeffrey R., Jeffrey B. Liebman and Lawrence F. Katz, Experimental Analysis of Neighborhood Effects", *Econometrica*, 75 (January 2007), 83-119.

Exemples of RCTs

Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2009. "The Miracle of Microfinance? Evidence from a Randomized Evaluation." *MIT Department of Economics.*

Behrman, Jere, Susan Parker, and Petra Todd. 2009. "Medium Term Impacts of the Oportunidades Conditional Cash Transfer Program on Rural Youth in Mexico" with, in Stephan Klasen and Felicitas Nowak-Lehmann, Eds., *Poverty, Inequality and Policy in Latin America, Cambridge* , MA: MIT Press.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Policies Have Displacement Effect? Evidence From a Clustered Random Experiment." *The Quarterly Journal of Economics* 128(2): 531-580.

Duflo, Esther, Michael Kremer, and Jonathan Robinson, "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya," Poverty Action Lab Working Paper
http://www.povertyactionlab.org/papers/99_Understanding_Technology_Adoption.pdf

Karlan, Dean and Jonathan Zinman, "Observing unobservables: Identifying Information Asymmetries with a Consumer Credit Field Experiment," (forthcoming in *Econometrica)*
http://karlan.yale.edu/p/OU_dec08_v1.pdf

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

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



Updated February 2010

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

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@coalition4evidence.org).

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

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

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

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

Checklist for overall study design

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-conducted randomized controlled trials conducted in various areas of social policy.[5]

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

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

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

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

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 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.

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

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

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

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

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

- Whether individuals (e.g., students) or groups (e.g., classrooms) were randomly assigned;
- Whether the sample was sorted into groups prior to randomization (i.e., "stratified," "blocked," or "paired"); and
- Whether the study intends its estimates of the intervention's effect to apply only to the sites (e.g., housing projects) in the study, or to be generalizable to a larger population.

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

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

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

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

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

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

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

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

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

References

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

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

> (a) The intervention may have sizeable "spillover" effects on individuals other than those who receive it.
>
> For example, if there is good reason to believe that a drug-abuse prevention program for youth in a public housing project may produce sizeable reductions in drug use not only among program participants, but also among their peers in the same housing project (through peer-influence), it is probably necessary to randomly assign whole housing projects to intervention and control groups to determine the program's effect. A study that only randomizes individual youth within a housing project to intervention versus control groups will underestimate the program's effect to the extent the program reduces drug use among both intervention and control-group students in the project.
>
> (b) The intervention is delivered to groups such as classrooms or schools (e.g., a classroom curriculum or schoolwide reform program), and the study seeks to distinguish the effect of the intervention from the effect of other group characteristics (e.g., quality of the classroom teacher).
>
> For example, in a study of a new classroom curriculum, classrooms in the sample will usually differ in two ways: (i) whether they use the new curriculum or not, and (ii) who is teaching the class. Therefore, if the study (for example) randomly assigns individual students to two classrooms that use the curriculum versus two classrooms that don't, the study will not be able to distinguish the effect of the curriculum from the effect of other classroom characteristics, such as the quality of the teacher. Such a study therefore probably needs to randomly assign whole classrooms and teachers (a sufficient sample of each) to intervention and control groups, to ensure that the two groups are equivalent not only in student characteristics but also in classroom and teacher characteristics.
>
> For similar reasons, a study of a schoolwide reform program will probably need to randomly assign whole schools to intervention and control groups, to ensure that the two groups are equivalent not only in student characteristics but also school characteristics (e.g., teacher quality, average class size).

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

[4] Resources that may be helpful in reviewing or conducting power analyses include: the William T. Grant Foundation's free consulting service in the design of group-randomized trials, at http://sitemaker.umich.edu/group- based/consultation_service; Steve Raudenbush et. al., *Optimal Design Software for Group Randomized Trials*, at http://sitemaker.umich.edu/group-based/optimal_design_software; Peter Z. Schochet, *Statistical Power for Random Assignment Evaluations of Education Programs* (http://www.mathematica- mpr.com/publications/PDFs/statisticalpower.pdf),

prepared for the U.S. Education Department's Institute of Education Sciences, June 22, 2005; and Howard S. Bloom, "Randomizing Groups to Evaluate Place-Based Programs," in *Learning More from Social Experiments: Evolving Analytical Approaches*, edited by Howard S. Bloom. New York: Russell Sage Foundation Publications, 2005, pp. 115-172.

[5] Here are illustrative examples of sample sizes from well-conducted randomized controlled trials in various areas of social policy: (i) 4,028 welfare applicants and recipients were randomized in a trial of Portland Oregon's Job Opportunities and Basic Skills Training Program (a welfare-to work program), to evaluate the program's effects on employment and earnings – see http://evidencebasedprograms.org/wordpress/?page_id=140; (ii) between 400 and 800 women were randomized in each of three trials of the Nurse-Family Partnership (a nurse home visitation program for low-income, pregnant women), to evaluate the program's effects on a range of maternal and child outcomes, such as child abuse and neglect, criminal arrests, and welfare dependency – see http://evidencebasedprograms.org/wordpress/?page_id=57; 206 9th graders were randomized in a trial of Check and Connect (a school dropout prevention program for at-risk students), to evaluate the program's effects on dropping out of school – see http://evidencebasedprograms.org/wordpress/?page_id=92; 56 schools containing nearly 6000 students were randomized in a trial of LifeSkills Training (a substance-abuse prevention program), to evaluate the program's effects on students' use of drugs, alcohol, and tobacco – see http://evidencebasedprograms.org/wordpress/?page_id=128.

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

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

Impact Evaluation Glossary

(Sources: 3ie and the World Bank)

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Null hypothesis
A null hypothesis is a hypothesis that might be falsified on the basis of observed data. The null hypothesis typically proposes a general or default position. In evaluation, the default position is

usually that there is no difference between the treatment and control group, or in other words, that the intervention has no impact on outcomes.

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

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

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

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

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

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

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

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

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

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

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

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

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

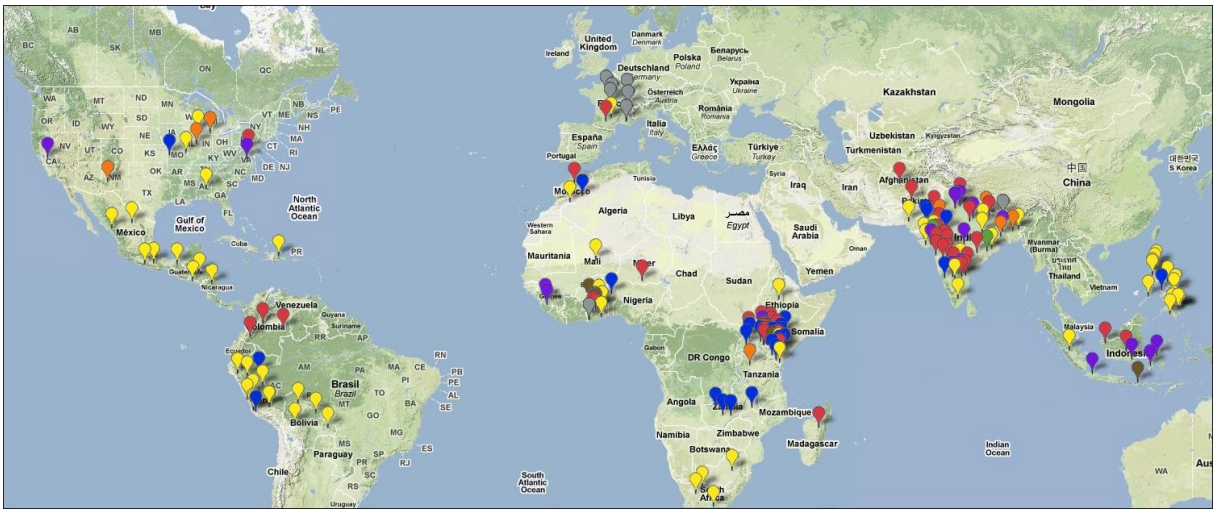
Stratified sample
Obtained by dividing the population of interest (sampling frame) into groups (for example, male and female), then by drawing a random sample within each group. A stratified sample is a probabilistic sample: every unit in each group (or strata) has the same probability of being drawn.

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

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

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

J-PAL & IPA Office Contacts



J-PAL projects all over the world

J-PAL Offices

J-PAL Global Office at MIT
30 Wadsworth St., E53-320
Cambridge, MA 02142 USA
Phone: +1 (617) 324-6566
Email: info@povertyactionlab.org
Website: www.povertyactionlab.org

J-PAL Africa Office at Southern Africa Labour & Development Research Unit (SALDRU)
University of Cape Town
Private Bag X3
Rondebosch 7701, SOUTH AFRICA
Phone: +27 21 650 5981
Email : jpalafrica@povertyactionlab.org
Website : www.povertyactionlab.org/africa

J-PAL Europe Office at Paris School of Economics
66bis avenue Jean Moulin
75014 Paris, FRANCE
Phone : +33 (0)1 71 19 40 70
Email : jpaleurope@povertyactionlab.org
Website : www.povertyactionlab.org/europe

J-PAL Latin America and Caribbean Office at Pontificia Universidad Católica de Chile
Instituto de Economía
Av. Vicuna Mackenna 4860
Santiago, CHILE
Phone: +(56-2) 354-1291
Email : jpallac@povertyactionlab.org
Website : www.povertyactionlab.org/LAC

J-PAL North America Office
30 Wadsworth St., E53-380
Cambridge, MA 02142 USA
Phone: +1 (617) 253 7109
Email: na-info@povertyactionlab.org
Website: www.povertyactionlab.org/north-america

J-PAL South Asia Office at the Institute for Financial Management and Research (IFMR)
IITM Research Park, A1, 10th Floor
Kanagam Road
Taramani, Chennai 600113, INDIA
Phone: +91 44 3247 50 56
Email : jpalsa@povertyactionlab.org
Website : www.povertyactionlab.org/south-asia

J-PAL Southeast Asia Office at Lembaga Penyelidikan Ekonomi dan Masyarakat (LPEM-UI)
Fakultas Ekonomi, Universitas Indonesia
Jl. Salemba Raya 4, Jakarta 10430, INDONESIA
Phone: +62 813 273 595 84
Email : jpalsea@povertyactionlab.org
Website : www.povertyactionlab.org/southeast-asia

IPA

Main Office:
Innovation for Poverty Action
101 Whitney Ave
New Haven CT 06510, USA
Phone: +1 203 772 2216
Email: contact@poverty-action.org
Website: www.poverty-action.org

For the other IPA Offices:
Please refer to www.poverty-action.org/about/contact

Notes

Notes

Notes

Notes

Notes

Notes

We thank Aude Guerrucci for her photographies. www.audeguerrucci.com