J-PAL

ABDUL LATIF JAMEEL

Poverty Action Lab

TRANSLATING RESEARCH INTO ACTION

[illegible]

Table of Contents

PROGRAMME

Executive Education Course in Evaluating Social Programmes, 16 – 19 May 2012

Accra, Ghana

	Wednesday **16 May 2012**	**Thursday** **17 May 2012**	**Friday** **18 May 2012**	**Saturday** **19 May 2012**
8:30 – 10:00	Welcoming Remarks Lecture 1*: What is Evaluation* Rebecca Thornton (University of Michigan)	Lecture 3: *Why Randomise* Kehinde Ajayi (Boston University)	Lecture 5: *Sampling and Sample Size* Paul Glewwe (University of Minnesota)	Lecture 7: *Project from Start to Finish* Kehinde Ajayi (Boston University)
10:30 – 12:30	Group Work Session 1 Introduction to group members Group work on case study 1: *School Monitoring Reform in Madagascar* Decision on group project (45min)	Group Work Session 3 Group work on case study 3: *Extra Teacher Programme* (60min) Group work on presentation (60min)	Group Work Session 5 Group exercise C on sample size estimation (60min) Group work on case study 4: *Evaluation Design* (60min)	Group Work Session 7 Group work to finalise presentations
	Lunch	Lunch	Lunch	Lunch
13:30 – 15:00	Lecture 2: *Measuring Impacts* Isaac Mbiti (Southern Methodist University)	Lecture 4: *How to Randomise* Paul Glewwe (University of Minnesota)	Lecture 6: *Threats and Analysis* TBD	**Group presentations** (each group: 15 min presentation, 15 min discussion)
15:30 – 17:00	Group Work Session 2 Group work on case study 2: *Learn to Read* (45min) Group work on presentation (45min)	Group Work Session 4 Group exercise A on random sampling (30min) and exercise B on mechanisms of randomisation (30min) Stats Primer (30 min)	Group Work Session 6 Group work on presentation (90min)	

Course Objectives

Our executive training programme is designed for people from a variety of backgrounds: managers and researchers from international development organisations, foundations, governments and non-governmental organisations from around the world, as well as trained economists looking to retool.

The course is a **full-time course**. It is important for participants to **attend all lectures and group work** in order to successfully complete the course and receive the certificate of completion.

Course Coverage

The following key questions and concepts will be covered:

- Why and when is a rigorous evaluation of social impact needed?
- The common pitfalls of evaluations, and how randomisation can help.
- The key components of a good randomised evaluation design.
- Alternative techniques for incorporating randomisation into project design.
- How do you determine the appropriate sample size, identify outcome measures, and manage data?
- Guarding against threats that may undermine the integrity of the results.
- Techniques for the analysis and interpretation of results.
- How to maximise policy impact and test external validity.

The programme will achieve these goals through a diverse set of integrated teaching methods. Expert researchers will provide both theoretical and example-based classes complemented by workshops where participants can apply key concepts to real world examples. By examining both successful and problematic evaluations, participants will better understand the significance of various specific details of randomised evaluations. Furthermore, the programme will offer extensive opportunities to apply these ideas ensuring that participants will leave with the knowledge, experience, and confidence necessary to conduct their own randomized evaluations.

J-PAL Lecturers

Kehinde Ajayi
Assistant Professor
Boston University

Kehinde Ajayi is an Assistant Professor at Boston University. Her research interests are in the areas of economic development and the economics of education, with a particular focus on school choice, educational mobility, and the effects of school quality on students' academic performance. She is currently conducting a set of studies relating to secondary education in Ghana.

Paul Glewwe
Professor of Applied Economics
University of University of Minnesota

Paul Glewwe is a Professor of Applied Economics at the University of Minnesota and the current Director of the Center for International Food and Agricultural Policy. His interests are economics of education, poverty and inequality in developing countries, and applied econometrics. His recent publications have appeared in the Handbook of the Economics of Education, Economic Development and Cultural Change, Journal of Development Economics, Journal of Economic Literature, Journal of Human Resources, Journal of Public Economics and World Bank Economic Review.

Isaac Mbiti
Assistant Professor
Southern Methodist University

Isaac Mbiti is an assistant professor in the Department of Economics of Southern Methodist University. His research interests are in economic development, labour economics and demography.

Rebecca Thornton
Assistant Professor of Economics
University of Michigan

Rebecca Thornton began her appointment as an assistant Professor at the University of Michigan economics department in 2008. Her research focuses on education and health as well as how individuals respond to financial incentives in these areas. She has worked on a randomized evaluation of a merit-based scholarship in Kenya. She is also working on randomized evaluations examining HIV testing and prevention and menstruation and education in Nepal.

List of Participants

#	Last Name	First Name	Organization	Country
1	Adamu-Issah	Madeez	UNICEF	Ghana
2	Adeola	Aminat	FATE Foundation	Nigeria
3	Amaniampong	Kwabena Adu	Council for Technical and Vocational Education and Training	Ghana
4	Asiegbor	Issac	Assessment Unit, CRDD	Ghana
5	Banashek	Sarah	USAID	Ghana
6	Brikorang	Fred	GES, Basic Education	Ghana
7	Brown	Victoria	Mango Tree	Uganda
8	Cohen	Lee	USAID	Mali
9	Coulibaly	Lazare	Institute for Popular Education	Mali
10	Ekuri	Emmanuel	African Population and Health Research Centre	Kenya
11	Hounkpodote	Hilaire	Laboratory for Research on Social Transformations	Senegal
12	Ilomu	Jessica	USAID	Uganda
13	Imoka	Chizoba	Unveiling Africa Foundation	Nigeria
14	Ka	Awa	ARED	Senegal
15	Kabaka	Stewart	MoPHS	Kenya
16	Kalanda	McKnight	Ministry of Education, Science and Technology	Malawi
17	Korr	Jacob	GES, Curriculum and Research Development	Ghana
18	Kotin	Veronica	EMIS	Ghana
19	Kwame	Agyeapong	Ministry of Education	Ghana
20	Lamptey	Elliot	Ministry of Education	Ghana
21	Lowe	Zev	Worldreader	Spain
22	Maiga	Eugenie	African Centre for Economic Transformation	Ghana
23	Mugerwa	Fred	Office of the Prime Minister	Uganda
24	Mukonka	Remmy	Ministry of Education, Science and Vocational Training	Zambia
25	Muvunyi	Emmanuel	Education Board	Rwanda
26	Oduro	Evelyn	GES, Teacher Education	Ghana
27	Ogle	Muktar	National Assessment Centre	Kenya
28	Oldmeadow	Emily	DFID	Nigeria
29	Otieno	Mary	Centre for Universal Education at Brookings	USA
30	Otim	Daniel	Mango Tree	Uganda
31	Otoo	Ernest	Ministry of Education	Ghana
32	Pealore	Dominic	Ministry of Education	Ghana
33	Perez	Marisol	USAID	Ghana
34	Rotich	Leah	Ministry of Education	Kenya
35	Ruto	Sara	Uwezo East Africa	Kenya
36	Salieu Kamara	Mohamed	Ministry of Education	Sierra

				Leone
37	Sarpong	Anthony	GES, Mathematics curriculum	Ghana
38	Thompson	Samuel	Council for Technical and Vocational Education and Training	Ghana
39	Traore	Bréhima	Oeuvre Malienne d'Aide à l'Enfance du Sahel	Mali
40	Umubyeyi	Marcienne	Self employed	Rwanda
41	Uneze	Eberechukwu	Centre for the Studies of the Economies of Africa	Nigeria
42	Weisenhorn	Nina	IRC	Congo

Group Assignment

Group 1
TA: Carolina Corral
1 Awa Ka
2 Bréhima Traore
3 Emmanuel Muvunyi
4 Hilaire Hounkpodote
5 Lazare Coulibaly
6 Marcienne Umubyeyi

Group 2
TA: Clare Hofmeyr
1 Eberechukwu Uneze
2 Emmanuel Ekuri
3 Eugenie Maiga
4 Fred Mugerwa
5 Mary Otieno
6 Stewart Kabaka

Group 3
TA: Elizabeth Schultz
1 Emily Oldmeadow
2 Jessica Ilomu
3 Lee Cohen
4 Marisol Perez
5 Nina Weisenhorn
6 Sarah Banashek

Group 4
TA: Amara Kallon
1 Aminat Adeola
2 Anthony Sarpong
3 Fred Brikorang
4 Issac Asiegbor
5 Samuel Thompson
6 Mohamed Salieu Kamara

Group 5
TA: Caitlin Tulloch
1 Dominic Pealore
2 Ernest Otoo
3 Kwabena Adu Amaniampong
4 Madeez Adamu-Issah
5 Veronica Kotin
6 Zev Lowe

Group 6
TA: Michael Polansky
1 Chizoba Imoka
2 Jacob Korr
3 Evelyn Oduro
4 McKnight Kalanda
5 Muktar Ogle
6 Remmy Mukonka

Group 7
TA: Pace Phillips
1 Agyeapong Kwame
2 Daniel Otim
3 Elliot Lamptey
4 Leah Rotich
5 Sara Ruto
6 Victoria Brown

Vocabulary

1. **Attrition**: the process of individuals joining in or dropping out of either the treatment or comparison group over the course of the study.

2. **Attrition Bias**: statistical bias which occurs when individuals systematically join in or drop out of either the treatment or the comparison group for reasons related to the treatment or outcomes.

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

4. **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.

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

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

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

8. **Equivalence**: groups are identical on all baseline characteristics, both observable and unobservable. Ensured by randomization.

9. **Externality**: an indirect cost or benefit incurred by individuals who did not directly receive the treatment. Also termed "spillover."

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

11. **Intention to Treat**: the measured impact of a program that includes all data from participants in the groups to which they were randomized, regardless of whether they actually received the treatment. Intention-to-treat analysis prevents bias caused by the loss of participants, which may disrupt the baseline equivalence established by randomization and which may reflect non-adherence to the protocol.

12. **Intra-cluster correlation coefficient**: a measure of the correlation between observations within a cluster; i.e. the level of correlation in drinl household.

13. **Level of Randomization:** the level of observation (ex. individual, household, school, village) at which treatment and comparison groups are randomly assigned.

14. **Partial Compliance**: individuals do not comply with their assignment (to treatment or comparison). Also termed "diffusion" or "contamination."

15. **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.

16. **Power**:the likelihood that, when the program has an effect, one will be able to distinguish the effect from zero given the sample size.

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

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

19. **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.

20. **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.

21. **Standardized Effect Size**: a standardized measure of the [expected] magnitude of the effect of a program.

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

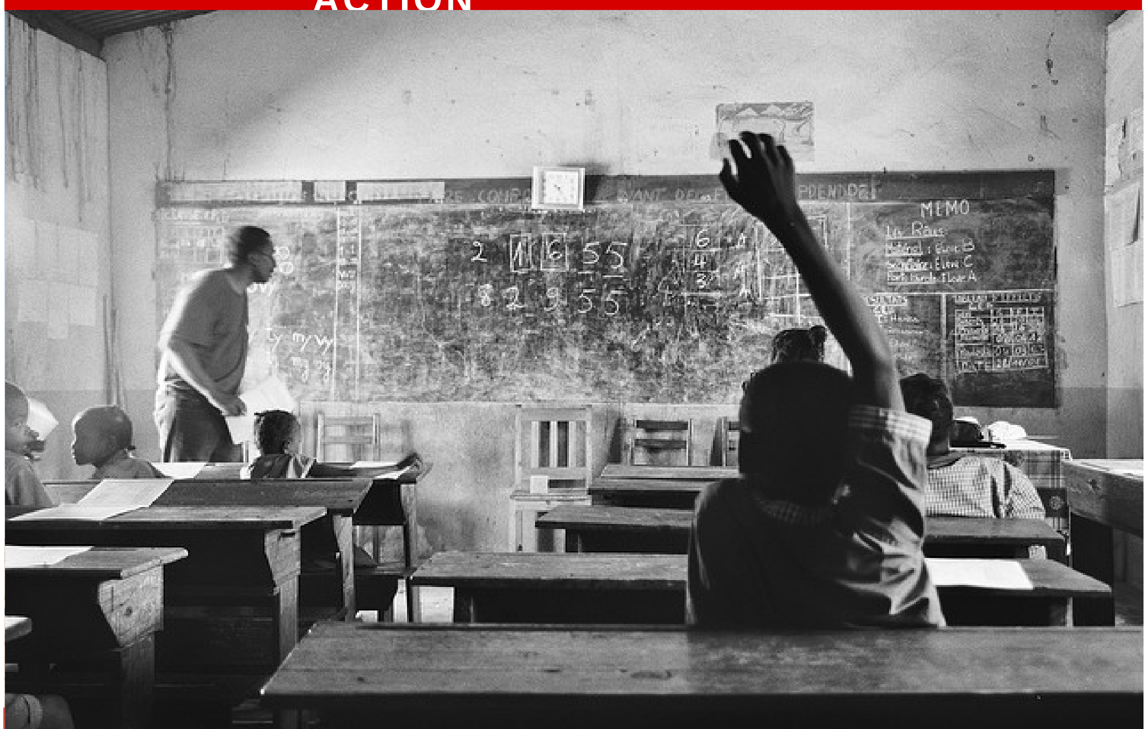
23. **Treatment Group**: in an experimental design, a randomly assigned group from the population that receives the intervention that is the subject of evaluation.

24. **Treatment on the Treated**: the measured impact of a program that includes only the data for participants who actually received the treatment.

CASE STUDY 1: REFORMING SCHOOL MONITORING

ABDUL LATIF JAMEEL
Poverty Action Lab

TRANSLATING RESEARCH INTO ACTION



Case 1: Reforming School Monitoring

Thinking about measurement and outcomes

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

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

Key Vocabulary

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

2. Indicators: metrics used to quantify and measure the needs that a program aims to address (needs assessment), how a program is implemented (process evaluation) and whether it affects specific short-term and long-term goals (impact evaluation).

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

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

Background

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

Madagascar School System Reforms: "Improving Outputs not Outcomes"

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

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

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

Discussion Topic 1

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

Problems remain….

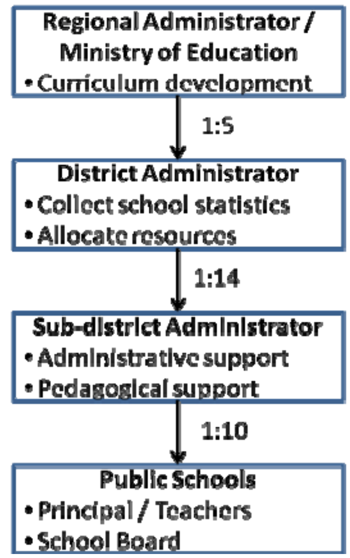
As the starting point of the study, researchers worked with the Ministry of Education to identify the remaining constraints in the schooling system. A survey conducted in 2005 revealed the following key problems:

1. **Teacher absenteeism:** At 10 percent, teacher absenteeism remains a significant problem. Only 8 percent of school directors monitor teacher attendance (either by taking daily attendance or tracking and posting a monthly summary of attendance), and more than 80 percent fail to report teacher absences to sub-district and district administrators.
2. **Communication with parents:** Communication between teachers and parents on student learning is often perfunctory, and student absenteeism is rarely communicated to parents.
3. **Teacher performance:** Essential pedagogical tasks are often neglected: only 15 percent of teachers consistently prepare daily and biweekly lessons plans while 20 percent do not prepare lesson plans at all. Student academic progress is also poorly monitored: results of tests and quizzes are rarely recorded and 25 percent of teachers do not prepare individual student report cards.

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

Intervention

Figure 1: Education System



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

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

1. **Top-Down:**

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

2. **Bottom-Up:**

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

Discussion Topic 2

1. Before setting up the RCT, researchers carefully analyzed the existing problem. Why do you think this is important as a starting point of an evaluation?

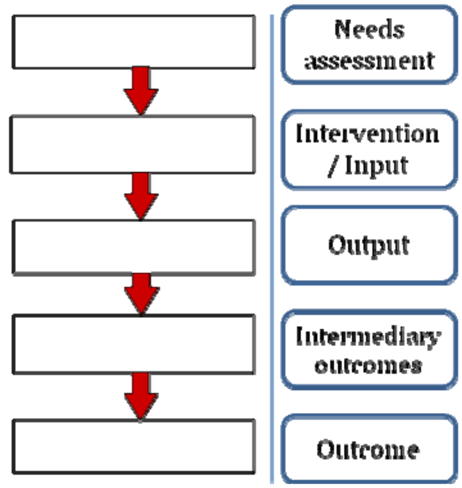
2. What are the intermediate and ultimate goals that this program hopes to achieve?

3. What is the key hypothesis being tested through this impact evaluation?

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

Theory of Change

Figure 2: Theory of Change



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

For example, a program or intervention is implemented to address a specific problem identified in the needs assessment (e.g. low literacy levels). The intervention (e.g. text books) may lead to outputs (e.g. students usage of textbooks) through which intermediary outcomes (e.g. reading skills) could be affected. These may lead to longer-term outcomes (e.g. drop-out rates, employment outcomes). An underlying assumption of this ToC is that students do not already have text books.

Discussion Topic 3

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

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

What data to collect? Data collection and measurement

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

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

Discussion Topic 4

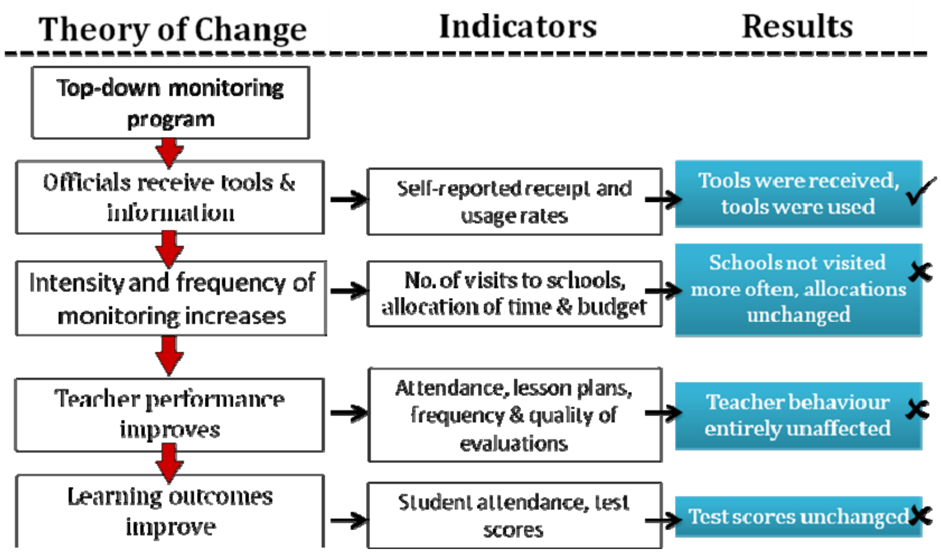
1. Which indicators would you measure at each step in the ToCs you drew up?

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

How to interpret the results

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

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



Discussion Topic 5

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

CASE STUDY 2: LEARN TO READ EVALUATIONS

ABDUL LATIF JAMEEL
Poverty Action Lab

TRANSLATING RESEARCH INTO ACTION

Case 2: Learn to Read Evaluations

Different methods of evaluation

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

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

Key Vocabulary

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

2. Treatment Group: in an experimental design, a randomly assigned group from the population that receives the intervention that is the subject of evaluation.

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

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

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

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

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

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

Why Learn to Read (L2R)?

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

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

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

Did the Learn to Read project work?

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

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

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

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

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

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

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

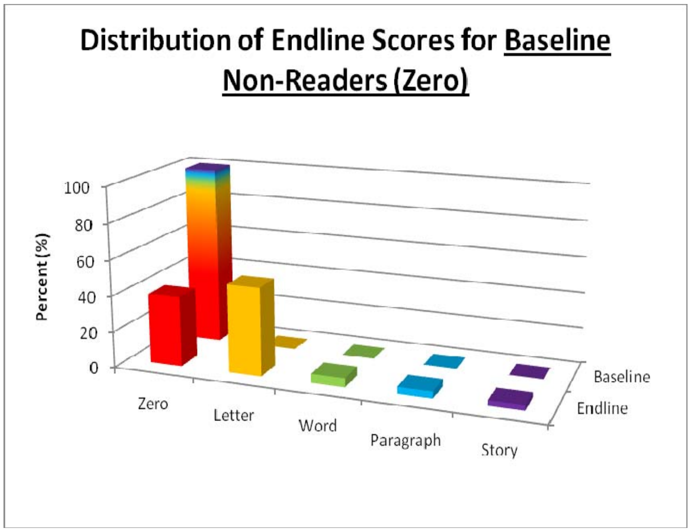
What comparison groups can we use? The following experts illustrate different methods of evaluating impact. (Refer to the table on the last page of the case for a list of different evaluation methods).

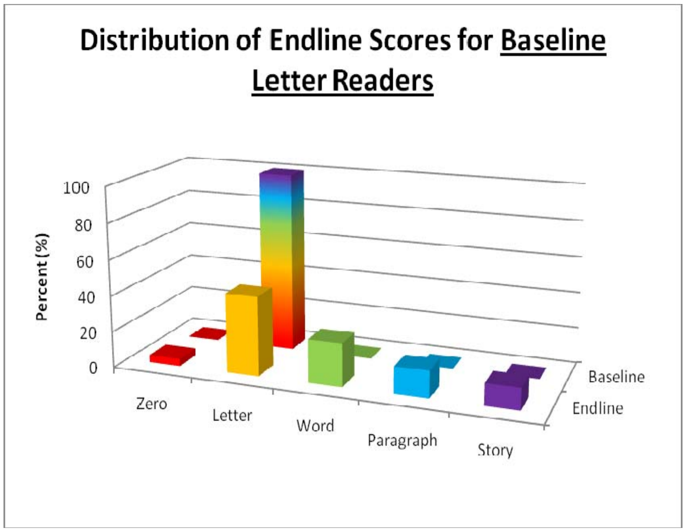
Estimating the impact of the Learn to Read project

Method 1

News Release: Read India helps children Learn to Read.

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





Figures: Show endline levels for L2R participants – organized by baseline levels of reading levels (left: zero level, right: letter level)

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

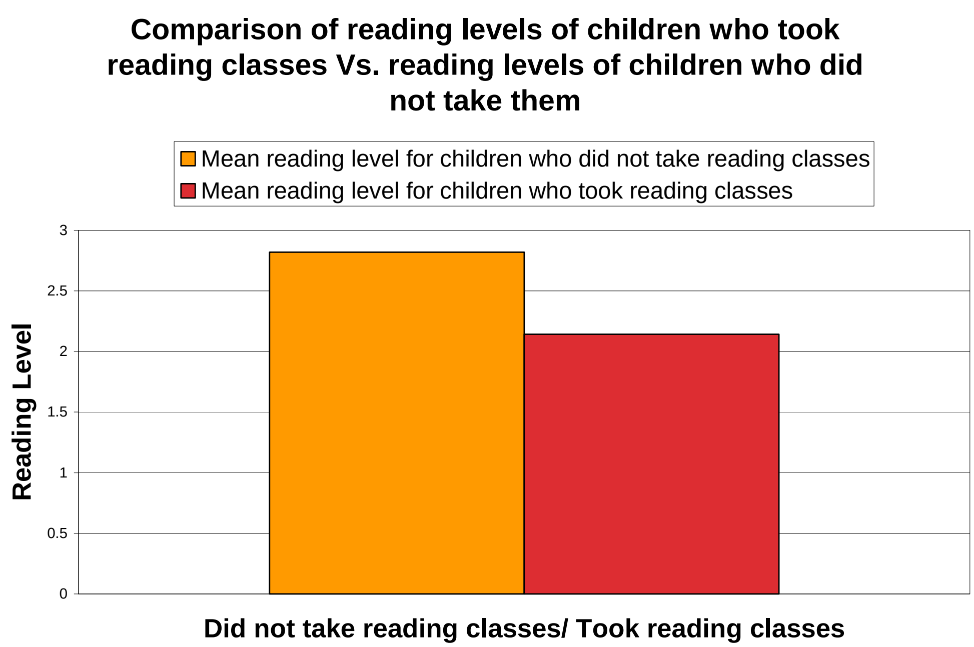
Discussion Topic 1

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

Method 2

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

Pratham has raised millions of dollars, expanding rapidly to cover all of India with its so-called "Learn-to-Read" program, but do its students actually learn to read? Recent evidence suggests otherwise. A team of evaluators from Education for All found that children who took the reading classes ended up with literacy levels significantly below those of their village counterparts. After one year of Pratham reading classes, Pratham students could only recognize words whereas those who steered clear of Pratham programs were able to read full paragraphs. If you have a dime to spare, and want to contribute to the education of India's illiterate children, you may think twice before throwing it into the fountain of Pratham's promises.



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

Discussion Topic 2

1. What type of evaluation is this opinion piece employing?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

Method 3

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

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

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

Discussion Topic 3

1. What type of evaluation is this letter using?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

Method 4

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

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

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

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

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

Table 1: Reading outcomes

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

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

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

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

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

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

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

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

Discussion Topic 4

1. What type of evaluation is this report utilizing?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

The table on the next two pages reviews the different impact evaluation methodologies.

	Methodology	**Description**	**Who is in the comparison group?**	**Required Assumptions**	**Required Data**
Quasi-Experimental Methods	**Pre-Post**	Measure how program participants improved (or changed) over time.	Program participants themselves—before participating in the program.	The program was the only factor influencing any changes in the measured outcome over time.	Before and after data for program participants.
	Simple Difference	Measure difference between program participants and non-participants after the program is completed.	Individuals who didn't participate in the program (for any reason), but for whom data were collected after the program.	Non-participants are identical to participants except for program participation, and were equally likely to enter program before it started.	After data for program participants and non-participants.
	Differences in Differences	Measure improvement (change) over time of program participants *relative to* the improvement (change) of non-participants.	Individuals who didn't participate in the program (for any reason), but for whom data were collected both before and after the program.	If the program didn't exist, the two groups would have had identical trajectories over this period.	Before and after data for both participants and non-participants.
	Multivariate Regression	Individuals who received treatment are compared with those who did not, and other factors that might explain differences in the outcomes are "controlled" for.	Individuals who didn't participate in the program (for any reason), but for whom data were collected both before and after the program. In this case data is not comprised of just indicators of outcomes, but other "explanatory" variables as well.	The factors that were *excluded* (because they are unobservable and/or have been not been measured) do not bias results because they are either uncorrelated with the outcome or do not differ between participants and non-participants.	Outcomes as well as "control variables" for both participants and non-participants.
	Statistical Matching	Individuals in control group are compared to similar individuals in experimental group.	Exact matching: For each participant, at least one non-participant who is identical *on selected characteristics*. Propensity score matching: non-participants who have a mix of characteristics which predict that they would be as likely to participate as participants.	The factors that were *excluded* (because they are unobservable and/or have been not been measured) do not bias results because they are either uncorrelated with the outcome or do not differ between participants and non-participants.	Outcomes as well as "variables for matching" for both participants and non-participants.
	Regression Discontinuity Design	Individuals are ranked based on specific, measureable criteria. There is some cutoff that determines whether an individual is eligible to participate. Participants are then compared to non-	Individuals who are close to the cutoff, but fall on the "wrong" side of that cutoff, and therefore do not get the program.	After controlling for the criteria (and other measures of choice), the remaining differences between individuals directly below and directly above the cut-off score are not statistically significant and will not bias the results. A necessary but	Outcomes as well as measures on criteria (and any other controls).

	Methodology	Description	Who is in the comparison group?	Required Assumptions	Required Data
		participants and the eligibility criterion is controlled for.		sufficient requirement for this to hold is that the cut-off criteria are strictly adhered to.	
	Instrumental Variables	Participation can be predicted by an incidental (almost random) factor, or "instrumental" variable, that is uncorrelated with the outcome, other than the fact that it predicts participation (and participation affects the outcome).	Individuals who, because of this close to random factor, are predicted not to participate and (possibly as a result) did not participate.	If it weren't for the instrumental variable's ability to predict participation, this "instrument" would otherwise have no effect on or be uncorrelated with the outcome.	Outcomes, the "instrument," and other control variables.
Experimental Method	**Randomized Evaluation**	Experimental method for measuring a causal relationship between two variables.	Participants are randomly assigned to the control groups.	Randomization "worked." That is, the two groups are statistically identical (on observed and unobserved factors).	Outcome data for control and experimental groups. Control variables can help absorb variance and improve "power".

CASE STUDY 3: EXTRA TEACHER PROGRAM

ABDUL LATIF JAMEEL
Poverty Action Lab

TRANSLATING RESEARCH INTO ACTION

Case 3: Extra Teacher Program

Designing an evaluation to answer three key education policy questions

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

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

Key Vocabulary

1. Level of Randomization: the level of observation (ex. individual, household, school, village) at which treatment and comparison groups are randomly assigned.

Over-crowded schools

Like many other developing countries, Kenya has recently made rapid progress toward the Millennium Development Goal of universal primary education. Largely due to the elimination of school fees in 2003, primary school enrollment rose nearly 30 percent, from 5.9 million to 7.6 million between 2002 and 2005.[2]

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

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

2. **Teacher absenteeism:** Further exacerbating the problem of pupil-teacher ratios, teacher absenteeism remains high, reaching nearly 20% in some areas of Kenya. There are typically no substitutes for absent teachers, so students simply mill around, go home or join another class, often of a different grade. Small schools, which are prevalent in rural areas of developing countries, may be closed entirely as a result of teacher absence. Families have to consider whether school will even be open when deciding whether or not to send their children to school. An obvious result is low student attendance—even on days when the school is open.

3. **Heterogeneous classes:** Classes in Kenya are also very heterogeneous with students varying widely in terms of school preparedness and support from home. Grouping students into classes sorted by ability (*tracking*, or *streaming*) is controversial among academics and policymakers. On one hand, if teachers find it easier to teach a homogeneous group of students, tracking could improve school effectiveness and test scores. Many argue, on the other hand, that if students learn in part from their peers, tracking could

[2] UNESCO. (2006). United Nations Education, Scientific and Cultural Organization. *Fact Book on Education for All*. Nairobi: UNESCO Publishing, 2006.

disadvantage low achieving students while benefiting high achieving students, thereby exacerbating inequality.

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

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

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

Contract Teachers: A possible solution?

Governments in several developing countries have responded to similar challenges by staffing unfilled teaching positions with locally-hired contract teachers who are not civil service employees. The four main characteristics of contract teachers are that they are:

1. Appointed on annual renewable contracts, with no guarantee of renewed employment (unlike regular civil service teachers)
2. Often less qualified than regular teachers and much less likely to have a formal teacher training certificate or degree
3. Paid lower salaries than those of regular teachers (typically less than a fifth of the salaries paid to regular teachers)
4. More likely to be from the local area where the school is located.

Are Contract Teachers Effective?

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

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

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

The Extra Teacher Program Randomized Evaluation

In January 2005, International Child Support (ICS) Africa initiated a two year program to examine the effect of contract teachers on education in Kenya. Under the program, ICS gave funds to 140 local school committees to hire one extra contract teacher to teach an additional first grade class. 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, because contract teachers tend to have less training and receive a lower monthly salary than their civil servant counterparts, there was concern about whether these teachers were sufficiently motivated, given their compensation, or qualified given their credentials.

The purpose of this intervention was to address the first three challenges: class size, teacher accountability, and heterogeneity of ability. The evaluation was designed to measure the impact of class-size reductions, the relative effectiveness of contract teachers, and how tracking by ability would impact both low and high-achieving students. What experimental designs could test the impact of this intervention on educational achievement? Which of these changes in the school landscape is primarily responsible for improved student performance?

Addressing Multiple Research Questions through Experimental Design

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

Discussion Topic 1: Testing the effectiveness of contract teachers

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

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

1. What is the effect of grouping students by ability on student performance?

2. What is the effect of smaller class sizes on student performance?

Discussion Topic 3: Addressing all questions with a single evaluation

1. Could a single evaluation explore all of these issues at once?

2. What randomization strategy could do so?

Exercise A: Random Sampling & The Law of Large Numbers (30 min) - Excel

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.

Your Group Leader has the data for this exercise.

1. Open the file "ExerciseA_SamplingDistributions_NEW.xlsm".

2. If prompted, select "Enable Macros".

3. Navigate to the "Randomize" worksheet, which allows you to choose a random sample of size "Sample Size" from the data contained in the "control" worksheet.

4. Enter "10" for "Sample Size and click the "Randomize" button. Observe the distribution of the various characteristics between Treatment, Control and Expected. With a sample size this small, the percentage difference from the expected average is quite high for reading scores. Click "Randomize" multiple times and observe how the distribution changes.

5. Now, try "50" for the sample size. What happens to the distributions? Randomize a few times and observe the percentage difference for the reading scores.

6. Increase the sample size to "500", "2000" and "10000", and repeat the observations from step 5. What can we say about larger sample sizes? How do they affect our Treatment and Control samples? Should the percentage difference between Treatment, Control and Expected always go down as we increase sample size?

Exercise B: Mechanics of Randomization (90min)

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 our list of all 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)

(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	

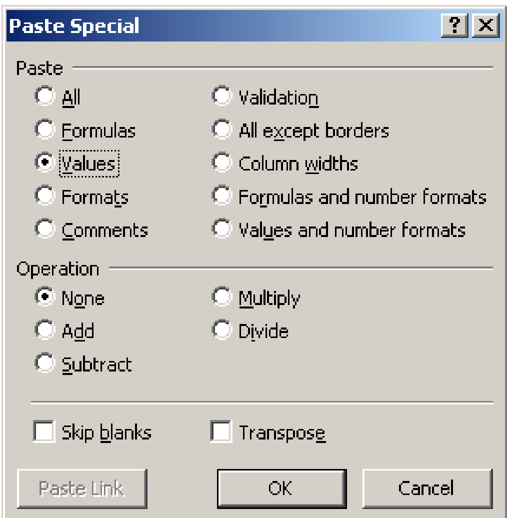
(3) Copy the cells in Colum C, then paste the values over the same cells

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.

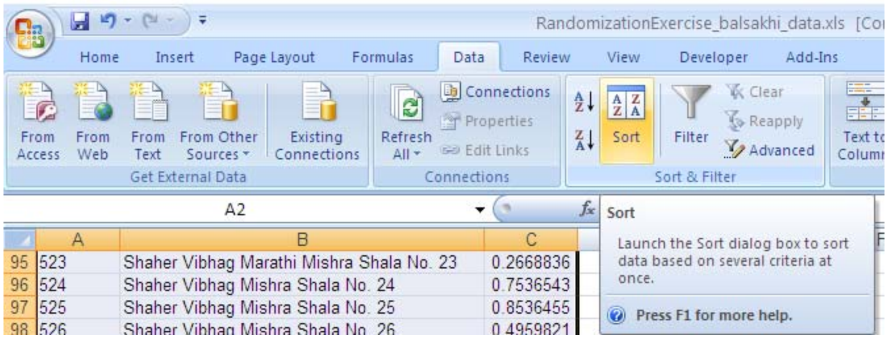
Then right click anywhere in that column and chose Paste Special. The "Paste Special window will appear. Click on "Values".



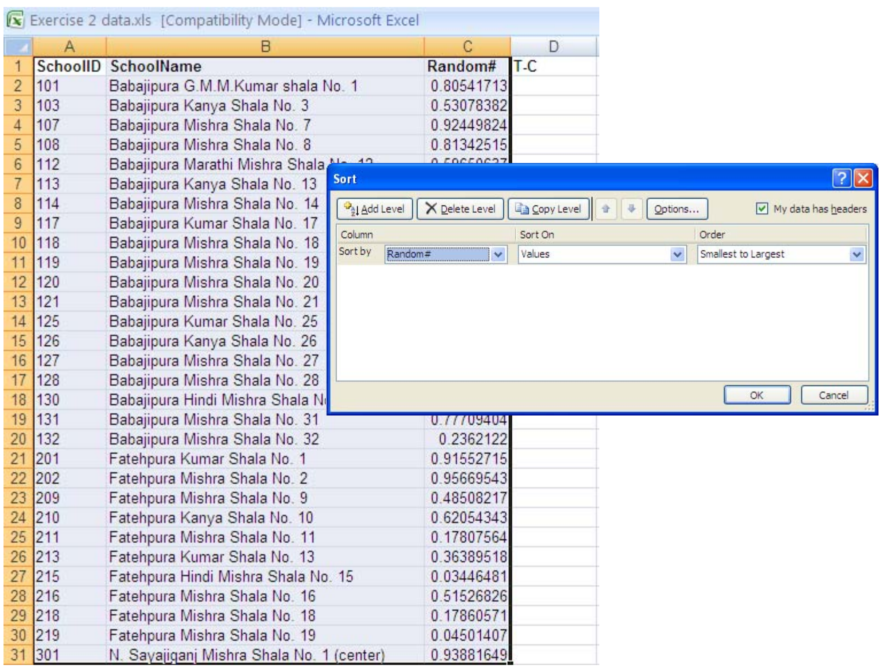
(4)

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

Highlight columns A, B, and C. In the data tab, 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		
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 N		
19	131	Babajipura Mishra Shala No. 31		
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.

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	
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

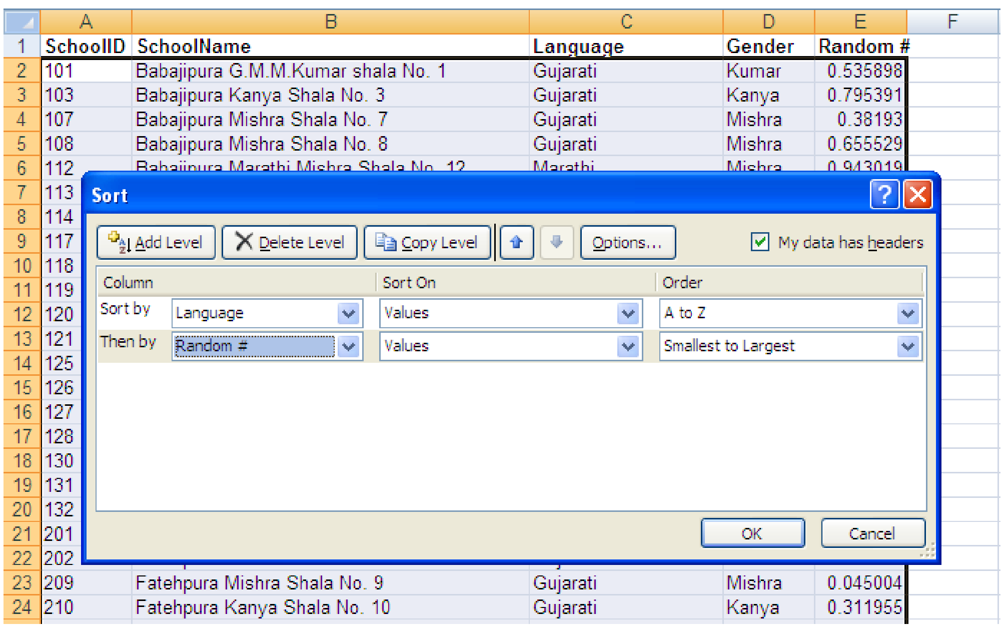
Stratification is the process of dividing a sample into groups, and then randomly assigning individuals within each group to the treatment and control. The reasons for doing this are rather technical. One reason for stratifying is that it ensures subgroups are balanced, making it easier to perform certain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

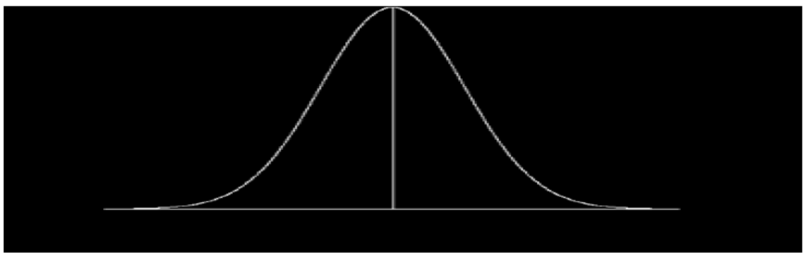
	A	B	C	D	E	F
100	132	Babajipura Mishra Shala No. 32	Gujarati	Mishra	0.8931975	C
101	615	Wadi Mishra Shala No. 15	Gujarati	Mishra	0.9142383	C
102	618	Wadi Kumar Shala No. 18	Gujarati	Kumar	0.9229356	C
103	408	Raopura Kanya Shala No. 8	Gujarati	Kanya	0.9285077	C
104	502	Shaher Vibhag Mishra Shala No. 2	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

Primer on Power Calculation (30 min)

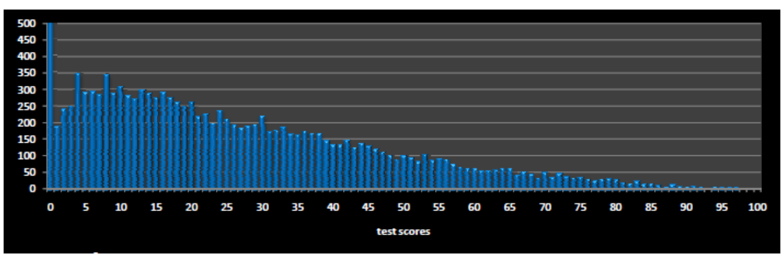
Sample means and the normal distribution

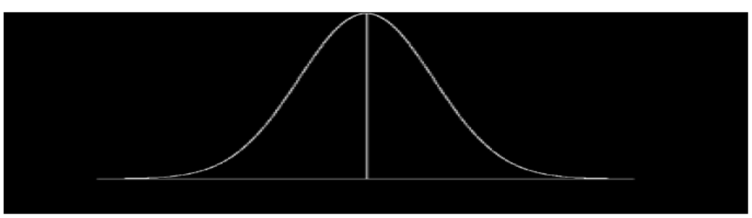
For statistics, we want to work with…

- A Normal Distribution…
- (a “Bell Curve”)



How do we get from here…

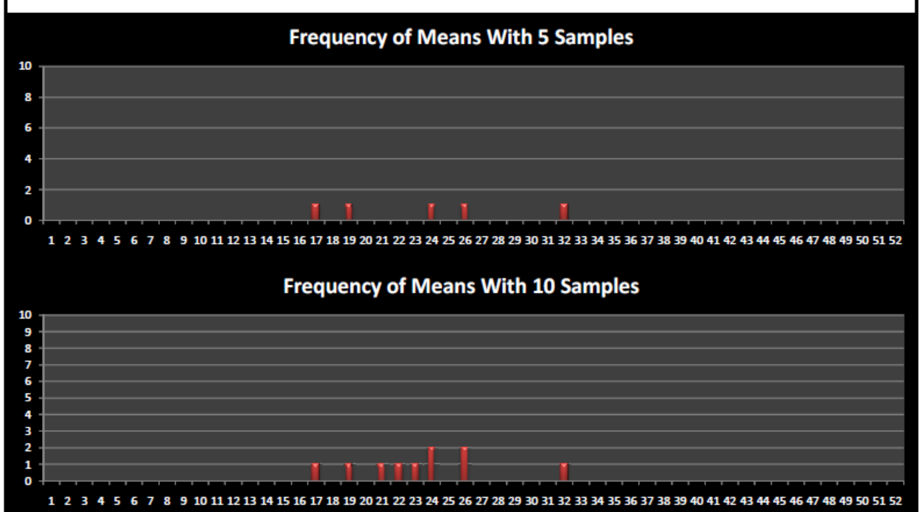


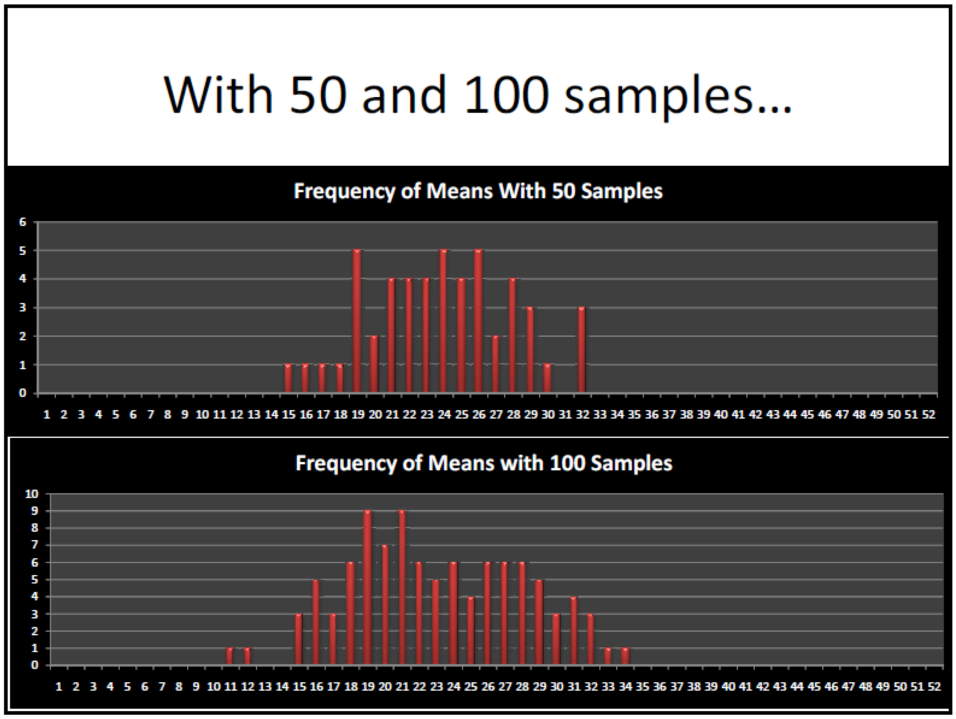


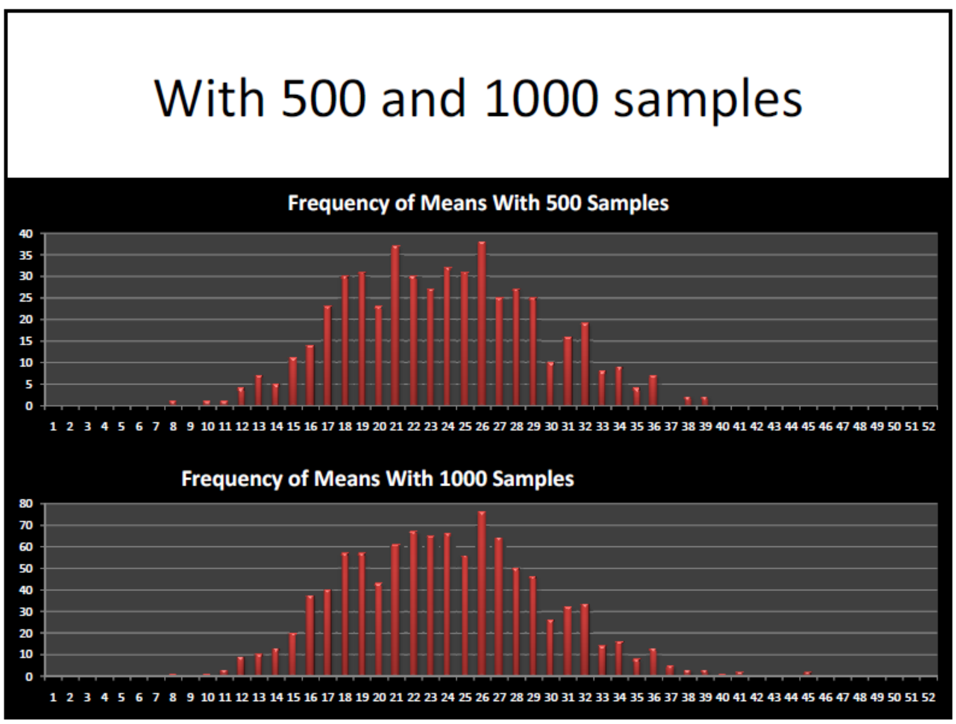
The normal distribution is

- A sampling distribution
- A distribution of "means" if we sampled the population an infinite number of times

Take 5 and 10 draws of 10…



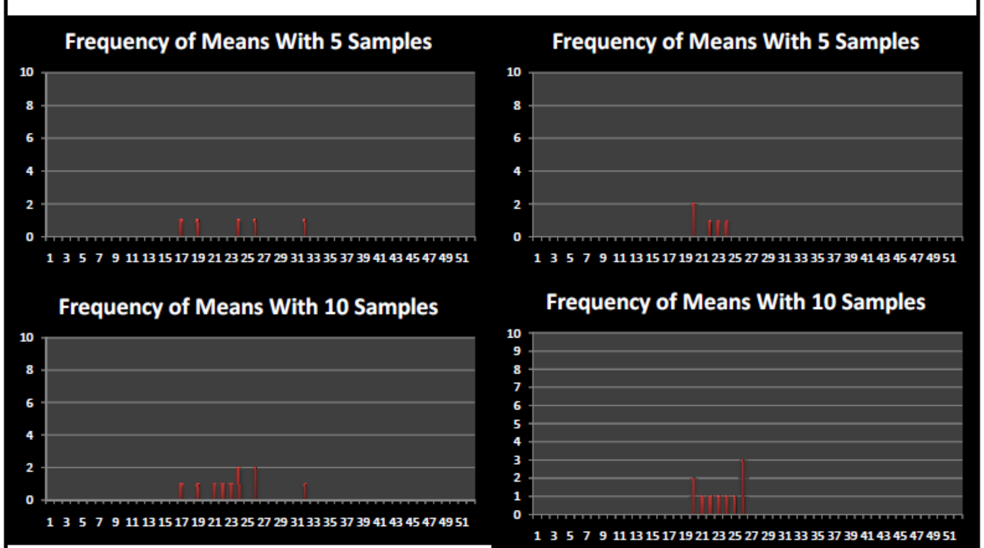

With 50 and 100 samples…
Frequency of Means With 50 Samples
Frequency of Means with 100 Samples

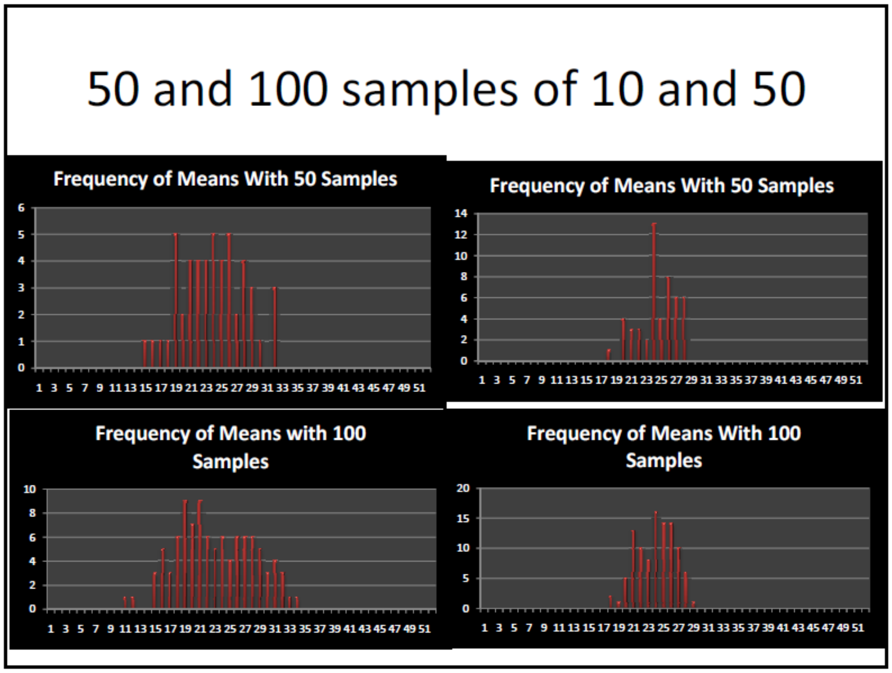

With 500 and 1000 samples
Frequency of Means With 500 Samples
Frequency of Means With 1000 Samples

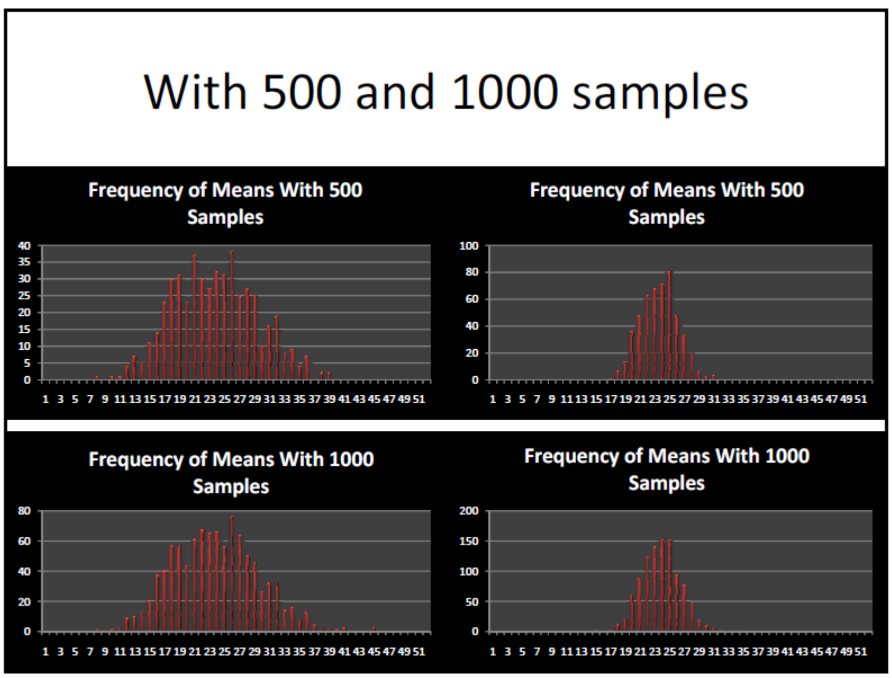
Size of Sample

- In our last example we drew samples of 10
  - (Sample size was 10)
- We observed the mean and plotted it on our graph
- After we repeated that 1000 times and looked at the "distribution of means," we saw it approached a bell curve
- What happens if we increase our sample size?

Take 5 and 10 draws of 10 and 50




50 and 100 samples of 10 and 50
Frequency of Means With 50 Samples
Frequency of Means With 50 Samples
Frequency of Means with 100 Samples
Frequency of Means With 100 Samples


With 500 and 1000 samples
Frequency of Means With 500 Samples
Frequency of Means With 500 Samples
Frequency of Means With 1000 Samples
Frequency of Means With 1000 Samples

Basic lessons:

- If we take enough samples from our data, the distribution of their means will eventually approach a normal distribution

  (Central Limit Theorem)
- More individuals per sample => quicker convergence to normal and mean

Group Exercise C: Power Calculation (30 min)

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.

Cluster randomized trials

The Extra Teacher Program (ETP) case study discussed the concept of cluster randomized trials. 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 extra teacher. 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 an extra year.

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 a 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 that 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 called Optimal Design developed by Steve Raudebush with funding from the William T. Grant Foundation. You can find additional resources on clustered designs on their web site.

Section 1: Using the OD Software

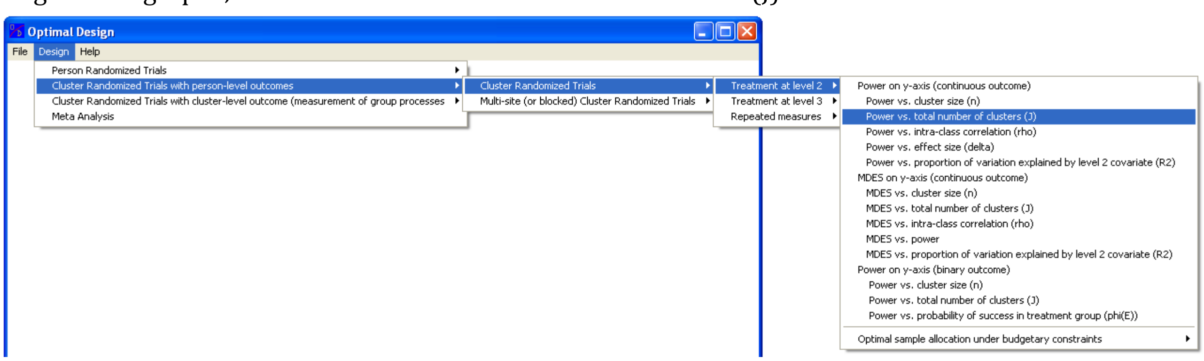
First download the OD software from the website (a software manual is also available):

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

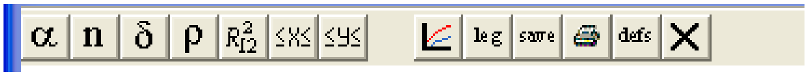
Notes on Optimal Design

- Menu
  - Personal Level Randomization (students)
  - Cluster with individual outcomes (schools randomized, look at individual test scores)
  - Cluster with cluster outcomes (school randomized, school level outcomes)
- Cluster with individual outcomes
  - Cluster Randomized Trial → assign clusters (schools, districts,etc) to C and T
  - Blocked Trial → stratification: ex. Stratify schools into similar pairs and randomize which are T and C
- Cluster Randomized Trial, Treatment at Level 2
  - Explain: five variables: rho, number of clusters, power, cluster size, and R^2… set three fixed (or at different levels for different lines) to get relationship between the other two on axis.

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



A new window will appear:



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

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

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

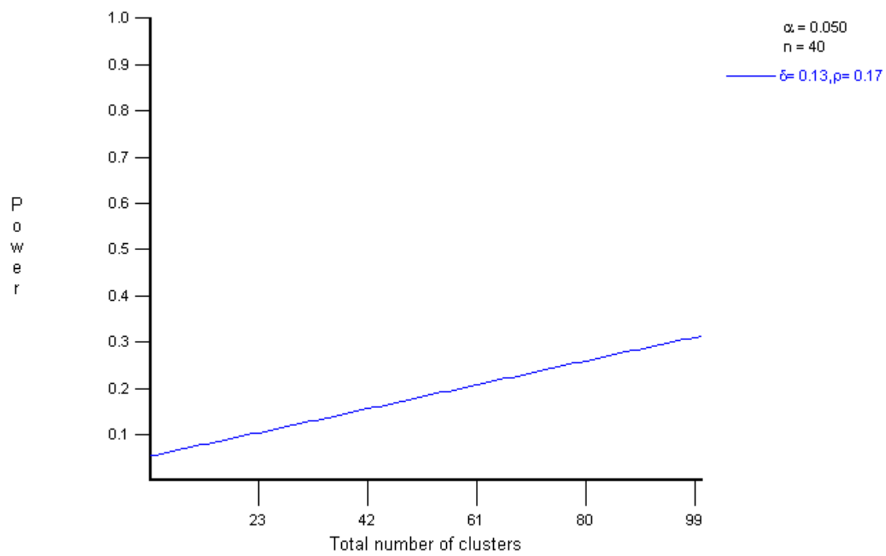
Now we have to determine δ (delta), the standard effect size (the effect size divided by the standard deviation of the variable of interest). Assume we are interested in detecting whether there is an increase of 10% in test scores. (Or more accurately, are uninterested in an increase of less than 10%.) 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 0.20 for delta (1). Change the value to 0.13, and change the value of delta (2) to empty. Select OK.

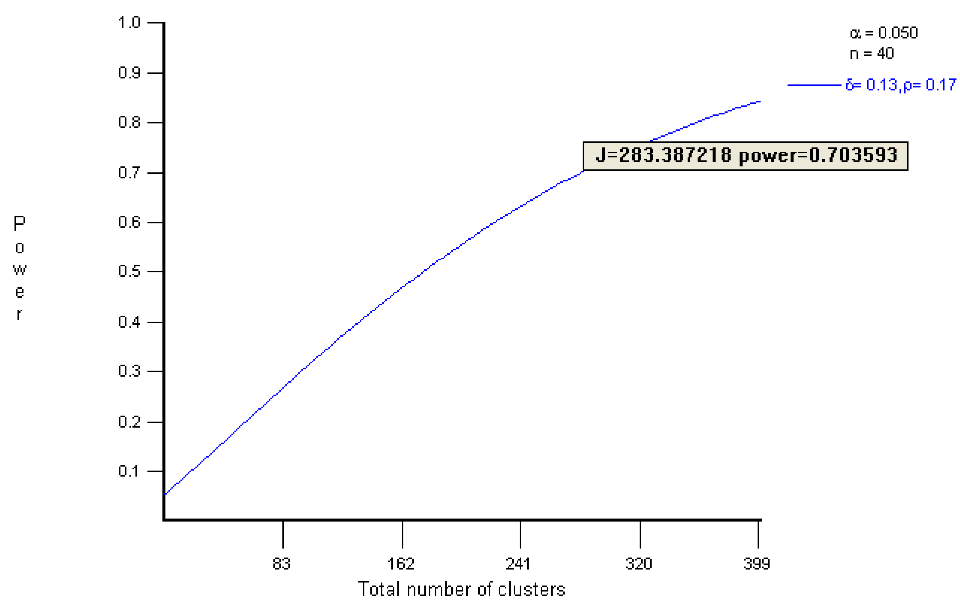
Finally we need to choose ***ρ*** (rho), which is the intra-cluster correlation. ***ρ*** tells us how strongly the outcomes are correlated for units within the same cluster. If students from the same school were clones (no variation) and all scored the same on the test, then ***ρ*** would equal 1. If,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

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

You should see a graph similar to the one below.



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



Exercise 1

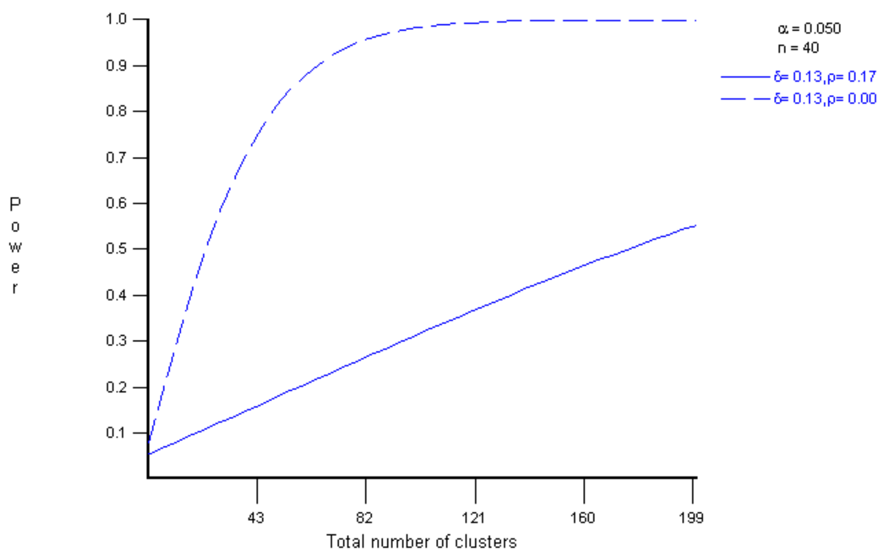
1. How many schools are needed to achieve 80% power? 90% power?

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

Exercise 2

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

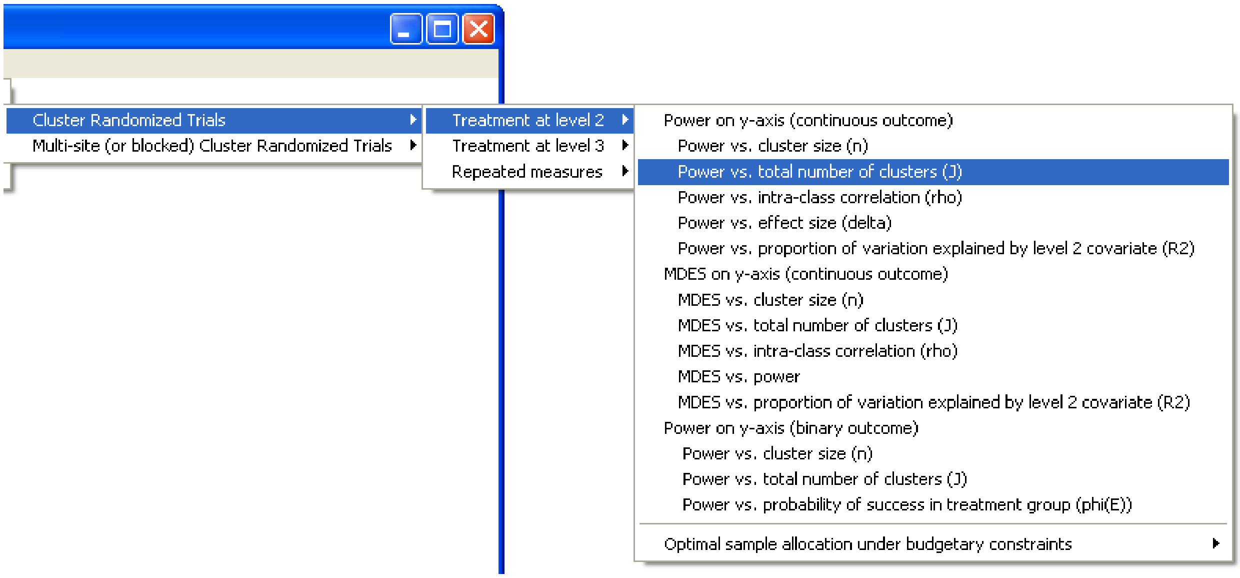
Finally, let's see how the Intraclass Correlation Coefficient (ρ) changes the power of a given sample. Leave rho(1) to be 0.17 but for comparison change rho(2) to 0.0. You should see a graph like the one below. The solid blue curve is the one with the parameters you've set. The blue dashed curve is there for comparison – to see how much power you would get from your sample if ρ were zero. Look carefully at the graph.



Exercise 3

1. 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

1. Try generating graphs for how power changes with the 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.

CASE STUDY 4: DEWORMING IN KENYA

ABDUL LATIF JAMEEL
Poverty Action Lab

TRANSLATING RESEARCH INTO ACTION

Case 4: Deworming in Kenya

Managing threats to experimental integrity

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

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

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 joining in or 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 join in or drop out of either the treatment or the comparison group for reasons related to the treatment or outcomes.

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 that includes all data from participants in the groups to which they were randomized, regardless of whether they actually received the treatment. Intention-to-treat analysis prevents bias caused by the loss of participants, which may disrupt the baseline equivalence established by randomization and which may reflect non-adherence to the protocol.

7. Treatment on the Treated: the measured impact of a program that includes only the data for participants who actually received the treatment.

8. Externality: an indirect cost or benefit incurred by individuals who did not directly receive the treatment. Also termed "spillover."

Worms - A common problem with a cheap solution

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

The symptoms associated with worm infections include listlessness, diarrhea, abdominal pain, and anemia. Beyond their effects on health and nutrition, heavy worm infections can impair children's physical and mental development and reduce their attendance and performance in school.

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

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

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

Primary School Deworming Program

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

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

Evaluation design — the experiment as planned

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

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

Threats to integrity of the planned experiment

Discussion Topic 1: Threats to experimental integrity

Randomization ensures that the groups are equivalent, and therefore comparable, at the beginning of the program. The impact is then estimated as the difference in the average outcome of the treatment group and the average outcome of the comparison group, 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?
3. Other than the program's direct and indirect impacts, what can happen over the course of the evaluation (after conducting the random assignment) to make the groups non-equivalent?
4. How does non-equivalence at the end threaten the integrity of the experiment?

Managing attrition - When the groups do not remain equivalent

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

Discussion Topic 2: Managing attrition

You are looking at the health effects of deworming. In particular you are looking at the worm load (severity of worm infection). Worm loads are scaled as follows:

- Heavy worm infections = 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 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 because of re-infection. Children who have a worm load of 3 only attend half the time and drop out of school if they are not treated. The number of children in each worm-load category is shown for both the pretest and posttest.

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

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

2. Besides worm load, the PSDP also looked at outcome measures such as school attendance rates and test scores.
   a) Would differential attrition (i.e. differences in drop-outs between treatment and comparison groups) bias either of these outcomes? How?
   b) Would the impacts on these final outcome measures be underestimated or overestimated?

3. In Case 1, 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 group does <u>not</u> get treated or the comparison group <u>does</u> get treated

Some people assigned to the treatment may in the end not actually get treated. In an after-school tutoring program, for example, some children assigned to receive tutoring may simply not show up for tutoring. And the others assigned to the comparison may obtain access to the treatment, either from the program or from another provider. Or comparison group children may get extra help from the teachers or acquire program materials and methods from their classmates. 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 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 none of the children from the poorest families have shoes and so they have worm loads of 3. Though their parents had not paid the school fees, the children were allowed to stay in school during the year. Parental consent was required for treatment, and to give consent, the parents had to come to the school and sign a consent form in the headmaster's office. However, because they had not paid school fees, the poorest 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	5,000	5000	5,000
2	5,000	5,000	0	5,000
1	5,000	5,000	10,000	5,000
Total children tested at school	15,000	15,000	15,000	15,000

1. a) Calculate the impact estimate based on the original group assignments.
   b) 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?

2. You are interested in learning the effect of treatment on those actually treated ("treatment on the treated" or ToT estimate). Five of your colleagues are passing by your desk; they all agree that you should calculate the effect of the treatment using only the 10,000 children who were treated.
   a) 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?

4. Another colleague suggests that you use the compliance rates, the proportion of people in each group that did or did not comply with their treatment assignment. You should divide the "intention to treat" estimate by the difference in the treatment ratios (i.e. proportions of each experimental group that received the treatment).
   a) Is this advice sound? Why or why not?

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

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

Discussion Topic 4: Managing spillovers

In the deworming program, randomization was at the school level. However, while all boys at a given treatment school were treated, only girls younger than thirteen received the deworming pill. This was due to the fact that the World Health Organization (WHO) had not tested (and thus not yet approved) the deworming pill for pregnant women. Because it was difficult to determine which girls were at risk of getting pregnant, the program decided to not administer the medication to any girl thirteen or older. (Postscript: since the deworming evaluation was implemented, the WHO has approved the deworming medication for pregnant women).

Thus at a given treatment school, there was a distinct group of students that was never treated but lived in very close proximity to a group that was treated.

Suppose protocol compliance is 100 percent: all boys and girls under thirteen in treatment schools get treated and all girls thirteen and over in treatment schools as well as all children in comparison schools do not get treated.

You can assume that due to proper randomization, the distribution of worm load across the three groups of students is equivalent between treatment and control schools prior to the intervention.

	Posttest					
	Treatment			**Comparison**		
Worm Load	All boys	Girls <13 yrs	Girls ≥ 13 yrs	All boys	Girls <13 yrs	Girls ≥ 13 yrs
3	0	0	0	5000	2000	2000
2	0	0	2000	5000	3000	3000
1	10000	5000	3000	0	0	0
Total children tested at school	20000			20000		

1. a) If there are any spillovers, where would you expect them to show up?
   b) Is it possible for you to capture these potential spillover effects? How?

2. a) What is the treatment effect for boys in treatment v. comparison schools?
   b) What is the treatment effect for girls under thirteen in treatment v. comparison schools?
   c) What is the direct treatment effect among those who were treated?
   d) What is the treatment effect for girls thirteen and older in treatment v. comparison schools?

e) What is the indirect treatment effect due to spillovers? f) What is the total program effect?

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

Group Presentation

Participants will form 4-6 person groups which will work through the design process for a randomised evaluation of a development project. Groups will be aided in this project by both the faculty and teaching assistants, with the work culminating in presentations at the end of the week.

The goal of the group presentation is to consolidate and apply the knowledge of the lectures and thereby ensure that participants leave with the knowledge, experience, and confidence necessary to conduct their own randomised evaluations. We encourage groups to work on projects that are relevant to participants' organisations.

All groups will present on Friday. The 15-minute presentation is followed by a 15-minute discussion led by J-PAL affiliates and staff. We provide groups with template slides for their presentation (see next page). While the groups do not need to follow this exactly, the presentation should have no more than 9 slides (including title slide, excluding appendix) and should include the following topics:

- Brief project background
- Theory of change
- Evaluation question
- Outcomes
- Evaluation design
- Data and sample size
- Potential validity threats and how to manage them
- Dissemination strategy of results

Please time yourself and do not exceed the allotted time. We have only a limited amount of time for these presentations and follow a strict timeline to be fair to all groups.

Group Presentation Template

Title

List your Team Members

You don't have to follow this exactly, this is just a guideline.

Background

- Talk briefly about general context, needs assessment, problem you want to solve.

Theory of Change

- Describe the specific intervention that you are evaluating.
- Talk about how it will solve part of the problem you described in the background.
- You may want to mention other causes of a problem that your intervention will not solve.
- (You can use the TOC template in the appendix.)

Evaluation Questions and Outcomes

- These should be directly linked to the TOC described above.
- What outcomes do you need to measure to test your research hypothesis?

Evaluation Design

- Unit of randomization, type of randomization (why did you choose these?)
- The actual randomization design- i.e. specific treatment group(s)

Data and Sample Size

- Outcomes
- Tell us where you will get the data – survey? Administrative?
- Power calcs
  - Justify where you got effect size and rho from, don't make it up.
  - You may need to do separate power calcs for separate outcomes.

Potential challenges

- Talk about threats (attrition, spillover, etc.) and how you want to manage them.
- You may need to revise your power calcs.

Results

- Why (and for whom) they would be useful.
- How would you disseminate them.

Appendix

Theory of change