



Executive Education Course Evaluating Social Programmes

University of Cape Town

18-22 January 2016



J-PAL

ABDUL LATIF JAMEEL POVERTY ACTION LAB
AFRICA

J-PAL Africa Executive Education



Agenda	2
Directions	3
Presenters	5
Participant Groups	6
Course Objectives	7
CASE STUDIES	
Case Study 1: Program Assessment - Reforming School Monitoring	8
Case Study 2: Learn to Read - Why Randomize	13
Case Study 3: Extra Teacher Program - How to Randomize	19
Case Study 4: Technoserve Coffee in Rwanda	22
EXERCISES	
Exercise A: Understanding Random Sampling and Law of Large Numbers	27
Exercise B: How to do Random Assignment using Ms Excel	28
Exercise C: How to do Power Calculations in Optimal Design Software	38
Guidelines for the Group Presentations	52
APPENDIX	
How to Read a Regression Table	54
Practical Tips for Cape Town	56

J-PAL Africa Executive Education Agenda

	Monday	Tuesday	Wednesday	Thursday	Friday	
8:30	<i>Tea and coffee</i>					
9:00	Registration	Measuring Impact <i>Martin Abel</i>	How to Randomise <i>Bruno Crepon</i>	Threats and Analysis <i>Bruno Crepon</i>	Cost Effectivness <i>Ashleigh Morrell</i>	
9:30					Locally Grounded and Globally Informed Policy advice <i>Emily Cupito</i>	
10:00					Introduction to J-PAL <i>Laura Poswell</i>	
10:15	Introduction to training <i>Ashleigh Morrell</i>	Break			Tea Break	
10:30	Baseline Assessment	Group work: Establishing a Research Question	Tea Break	Tea Break	Endline Assessment	
11:00	Tea Break	Tea break	Sample Size I <i>Jeremy Magruder</i>			
11:45	What is an Evaluation? <i>Jeremy Magruder</i>	Why Randomise <i>Martin Abel</i>	Sample size II <i>Jeremy Magruder</i>	Case Study 4: Threats and Analysis	Group work: Threats and Analysis	
12:00						
12:30						
13:00	Lunch	Lunch and measurement fair	Lunch: Make connections day	Lunch	Lunch and Feedback session	
13:30						
14:00						
14:30	Case study 1 : Reforming Schooling in Madagascar	Group work: Theory of Change and Indicators	Case study 3: How to Randomise: Extra Teacher programme	Project from Start to Finish <i>Bruno Crepon</i>	Participant presentations	
15:00			Tea Break			Tea Break
15:30			Tea break			Tea Break
15:45		Tea Break	Exercise C: Power Calculations	Group work: Randomisation Design and Power Calculations	Tea Break	
16:00	Case study 2: Why Randomise	Exercise A&B : Randomisation Mechanics	Participant Presentations			
16:30						
17:00						
17:30			Braai			
18:00						
18:30						

Cape Town – University of Cape Town



Directions to UCT Middle Campus from the airport

To reach the university from the airport proceed on the N2 towards Cape Town and take the Muizenberg (M3) off-ramp. Continue until you reach and turn off at the Woolsack Drive / University of Cape Town off ramp. Go straight at the traffic lights on Woolsack Drive and enter middle campus. Follow Cross Campus Road until you come to a stop sign. Take a left and after 100m you see the parking lot for the All Africa House and New Economics Building on the left side (**K3** on map on next page).

Directions to UCT Middle Campus from down town Cape Town

UCT's Middle Campus (Groote Schuur Campus) is situated on the slopes of Devil's Peak in the suburb of Rondebosch. To reach the middle campus from the city, drive along De Waal Drive or Eastern Boulevards, passing Groote Schuur Hospital on the way. Just past the hospital the road forks. Take the right-hand fork (M3 to Muizenberg). Just beyond Mostert's Mill (windmill) on your left, take the Woolsack Drive / University of Cape Town turn-off. Go straight at the traffic lights on Woolsack Drive and enter middle campus. Follow the road until you come to a stop sign. Take a left and after 100m you will see the parking lot for the All Africa House and New Economics Building on the left side (**K3** on map on next page)

UCT Middle Campus: New Economics Building is in Cell K3 below:



Presenters



Martin Abel

Martin Abel is a PhD Candidate in Public Policy at the Harvard Kennedy School. In his previous capacity as a research manager at J-PAL Africa, he worked on issues of financial literacy and youth employment in cooperation with the South African government. His research interests are development economics, behavioral economics and labor economics. He holds a Master in Public Administration in International Development (MPA/ID) from the Harvard Kennedy School and a Diploma in Economics from the University of Jena.



Bruno Crepon

Bruno Crépon is a researcher at Centre de Recherche en Économie et Statistique (CREST) and an Associate Professor at ENSAE and École Polytechnique. The focus of his research is on policy evaluation with special attention to labor market policies.



Jeremy Magruder

Jeremy Magruder is an associate professor in the Department of Agricultural and Resource Economics, University of California, Berkeley. His research interests include unemployment in South Africa, job networks in India, and HIV/AIDS in Malawi.



Emily Cupito

Emily Cupito works as a Policy Manager for J-PAL Africa at the University of Cape Town. She leads outreach to practitioners and policymakers across the continent. She helps policymakers interpret rigorous research results and think strategically about how these results can be translated into effective programs. Prior to her work at J-PAL, Emily spent more than two years working in Uganda with Innovations for Poverty Action, where she supported financial inclusion research by leading dissemination efforts, developing new projects, and working to build the capacity of researchers in Africa and South Asia. She previously worked as a Presidential Management Fellow with the US Federal Government. Emily received a Master's in Public Policy from Duke University and a Bachelor's in Economics and Public Policy from the University of North Carolina at Chapel Hill.

Participant Groups

Âurea and Ashleigh	Room : 4.49 (Boardroom)
Shukri	Mohamed
Marylène	Wamukoya
Damien	Kirchhoffer
Leala	Ruangtragool
Nathene	Morley
Moroesi	Maketha

Kathryn	Room: 3.13
Emily	Todd
Annlin	Matabane
Mikateko	Mbambo
Ben	Maiyo
Herbert	Ntuli
Nicola	Jowell
Anne	Morgenroth

Emily and Chris	Room: 5.06
Douglas	Montoya
Alejandra	Campero
David	Rinnert
Wisdom	Dube
Abdul	Muwanika
Yousuf	Caires

Olawunmi	Room: 4.07
Ugo	Pica-Ciamarra
Nuri	Niyazi
Srinivasulu	Rajendran
Thomas	Harris
Danielle	Dunn

Emmanuel	Room: 3.06 (Boardroom)
Andreea	Minea
Alexandra	Ames
Solange	Rosa
Hazel Laura	Mugo
Carrie	Mashek
Kathryn	Huston

Course Objectives

This impact evaluation training course is designed for people from a variety of backgrounds: managers and researchers from international development organizations, foundations, governments and non-governmental organizations from around the world, as well as trained economists looking to retool.

The course is scheduled to run 5 full days. 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.

Key Questions

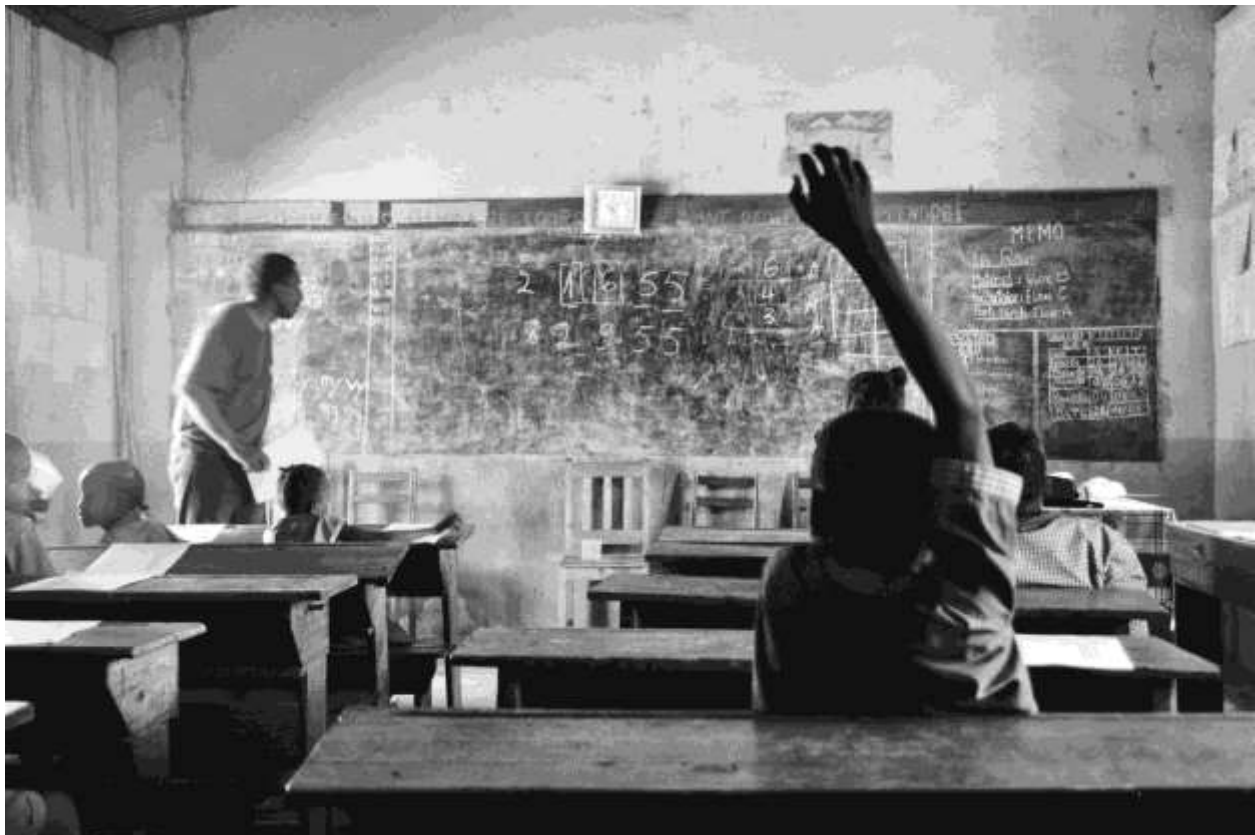
The following key questions and concepts will be covered:

- Why and when is a rigorous evaluation of social impact needed?
- The common pitfalls of evaluations, and why does randomization help.
- What are the key components of a good randomized evaluation design?
- Alternative techniques for incorporating randomization into project design.
- How do you determine the appropriate sample size, measure outcomes, and manage data.
- Guarding against threats that may undermine the integrity of the results.

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 engage with research using randomised evaluations.

CASE STUDY 1: REFORMING SCHOOL MONITORING

Program Theory and Measuring 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 paper.

KEY VOCABULARY

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

Indicators: metrics used to quantify and measure specific short-term and long-term effects of a program.

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

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

INTRODUCTION

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

Madagascar school system reforms

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

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 evaluate¹:

1. Top-Down

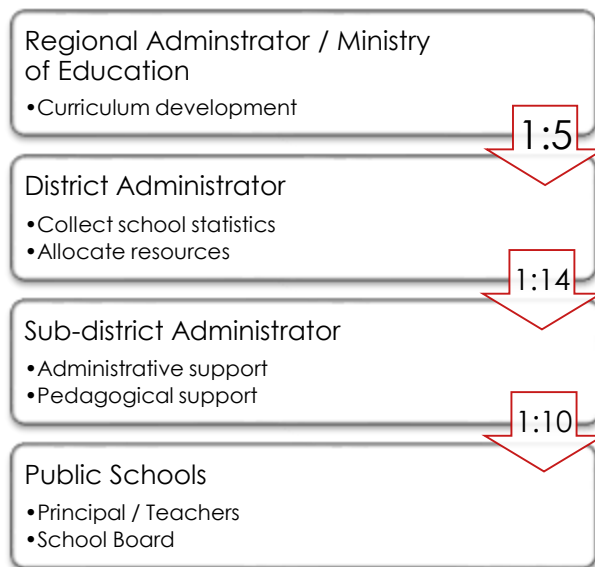
Operational tools and guidebooks which outline their responsibilities are given to the relevant administrators. During a meeting, administrators are trained on how to carry out their tasks, and their performance criteria are clarified. This is followed up by regular monitoring of their performance, which 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.

FIGURE 1: EDUCATION SYSTEM



Discussion Topic 2

Intermediate and final outcomes

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?

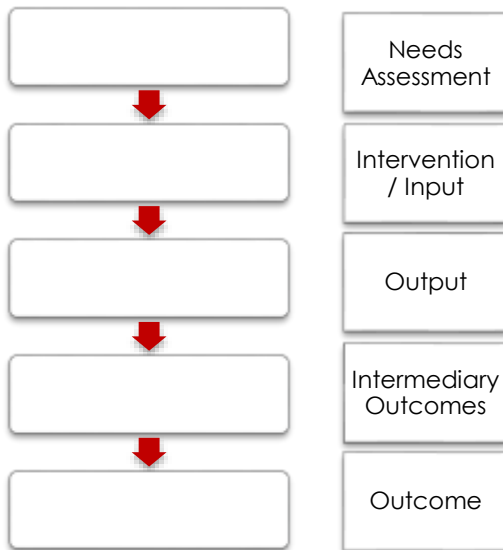
THEORY OF CHANGE

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.

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.

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.

FIGURE 2: THEORY OF CHANGE



Discussion Topic 3

Theory of change

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

Measuring outcomes and indicators

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

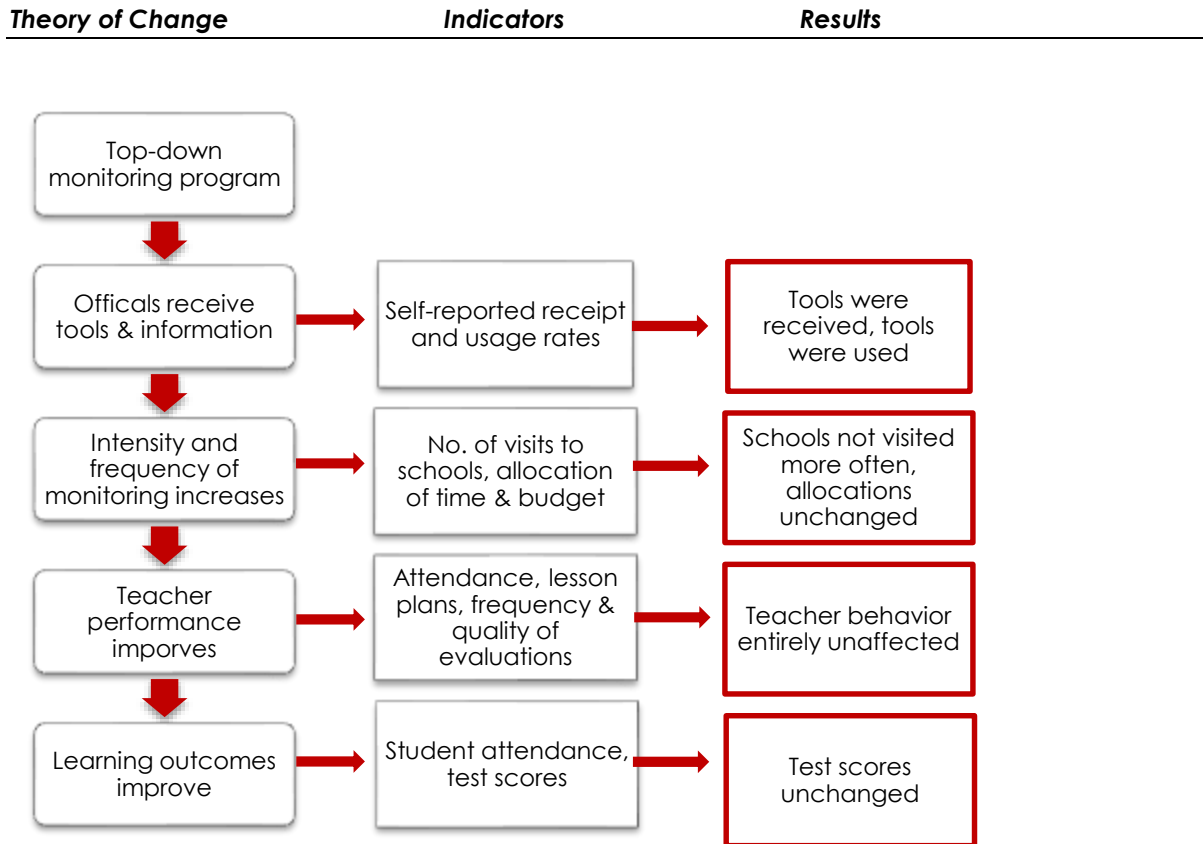
The evaluation found that the bottom-up approach led to successful results. Attendance at meetings between teachers and community members was high, and although communication between teachers and parents did not change, teachers improved the quality of teaching as shown by an increase in lesson plans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

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

Discussion Topic 5

Interpreting the results

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

Why Randomize?



This case study is based on “Remedying Education: Evidence from Two Randomized Control Evaluations in India” by Abhijit Banerjee, Shawn Cole, Esther Duflo, and Leigh Linden.

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

Key Vocabulary

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

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

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

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

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

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

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

Introduction

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

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

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

Did this program work? How would you measure the impact?

Did the Learn to Read Project work?

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

In general, to ask if a program works is to ask if the program achieves its goal of changing certain outcomes for its participants, and ensure that those changes are not caused by some other factors or events happening at the same time. To show that the program causes the observed changes, we need to simultaneously show that if the program had not been implemented, the observed changes would not have occurred (or would be different). But how do we know what would have happened? If the program happened, it happened. Measuring what would have happened in the absence of the program 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 our case, the intention of the program is to “improve children’s reading levels” and the reading level is the outcome measure. So, when we ask if the Learn to Read project worked, we are asking if it improved children’s reading levels. The impact is the difference between reading levels after the children have taken the reading classes and what their reading level would have been if the reading classes had never existed.

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

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

Estimating the impact of the Learn to Read project

METHOD 1:

News Release: Read India helps children Learn to Read.

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

FIGURE 1

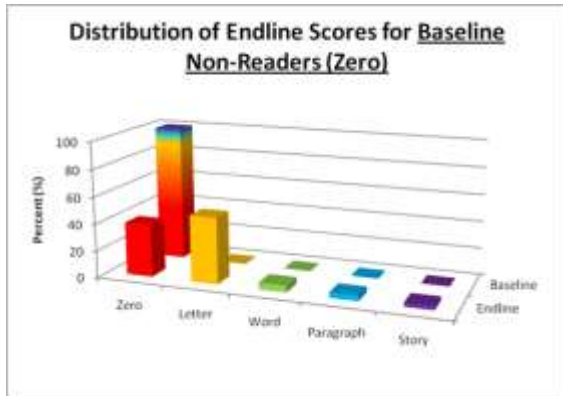
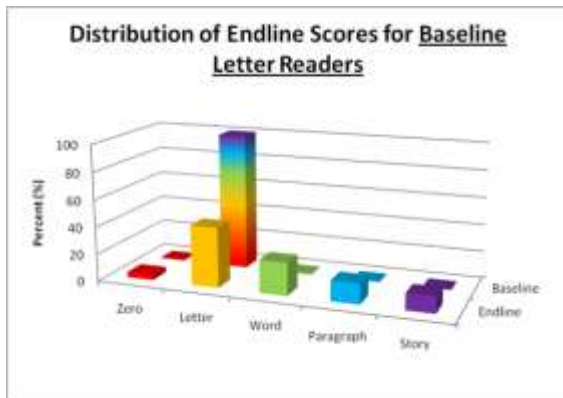


FIGURE 2



Just before the program started, half these children could not recognize Hindi word. But after spending just a few months in Pratham reading classes, more than half improved by at least one reading level, with a significant number capable of recognizing words and several able to read full paragraphs and stories! *On average, the literacy measure of these students improved by nearly one full reading level during this period.*

DISCUSSION TOPIC 1

Identifying evaluation

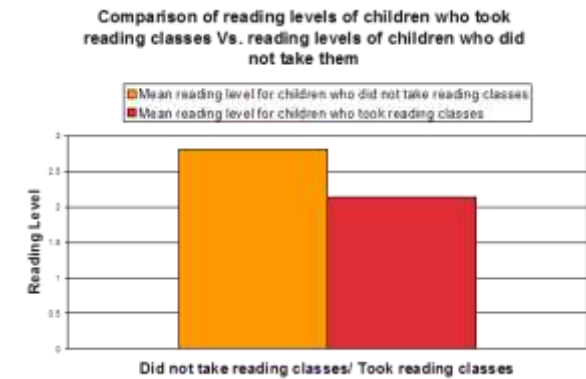
1. What type of evaluation does this news release imply?
2. What represents the counterfactual?
3. What are the problems with this type of evaluation?

METHOD 2:

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

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

FIGURE 3



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

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

DISCUSSION TOPIC 2

Identifying evaluation

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

METHOD 3:

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

There have been several unfair reports in the press concerning programs implemented by the NGO Pratham. A recent article by a former Education for All bureaucrat claims that Pratham is actually hurting the children it recruits into its ‘Learn-to-Read’ camps. However, the EFA analysis uses the wrong metric to measure impact. It compares the reading *levels* of Pratham students with other children in the village—not taking into account the fact that Pratham targets those whose literacy levels are particularly poor at the beginning. If Pratham simply recruited the most literate children into their programs, and compared them to their poorer counterparts, they could claim success without conducting a single class. But Pratham does not do this. And realistically, Pratham does not expect its illiterate children to overtake the stronger students in the village. It simply tries to initiate improvement over the current state.

Therefore the metric should be *improvement* in reading levels—not the final level. When we repeated EFA’s analysis using the more-appropriate outcome measure, the Pratham kids improved at twice the rate of the non-Pratham kids (0.6 reading level increase compared to 0.3). This difference is statistically very significant.

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

DISCUSSION TOPIC 3

Identifying evaluation

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

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

How to Randomize?



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

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

KEY VOCABULARY

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

INTRODUCTION

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

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

OVERCROWDED SCHOOLS

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

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

1. **Large class sizes:** Due to budget constraints, the rise in primary school enrollment has not been matched by proportional increases in the number of teachers. (Teacher salaries already account for the largest component of educational spending.) The

result has been very large class sizes, particularly in lower grades. In a sample of schools in Western Kenya, for example, the average first grade class in 2005 had 83 students. This is concerning because it is believed that small classes are most important for the youngest students, who are still acclimating to the school environment. The Kenyan National Union of Teachers estimates that the country needs an additional 60,000 primary school teachers in addition to the existing 175,000 in order to reach all primary students and decrease class sizes.

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

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

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

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

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

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

WHAT ARE CONTRACT TEACHERS?

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

ARE CONTRACT TEACHERS EFFECTIVE?

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

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

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

THE EXTRA TEACHER PROGRAM RANDOMIZED EVALUATION

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

ADDRESSING MULTIPLE RESEARCH QUESTIONS THROUGH EXPERIMENTAL DESIGN

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

Discussion Topic 1

Testing the effectiveness of contract teachers

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

Discussion Topic 2

Looking at more general approaches to improving education

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

Discussion Topic 3

Addressing all questions with a single evaluation

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

CASE STUDY 4: TECHNOSERVE COFFEE IN RWANDA

Addressing threats to experimental integrity



This case study is based on a current study by Esther Duflo, Tavneet Suri Daniel Keniston.

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

KEY VOCABULARY

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.

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

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

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

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

Intention to Treat: the measured impact of a program comparing study (treatment versus control) groups, regardless of whether they actually received the treatment.

Treatment on Treated: the measured impact of a program on those who actually received or took up the treatment.

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

INTRODUCTION

In 2010, the Technoserve (TNS) Coffee Initiative partnered with J-PAL researchers to conduct a randomized evaluation on their coffee agronomy-training program in Nyarubaka sector in southern Rwanda. Technoserve carried out their regular recruitment sign-up processes across all 27 villages in the sector and registered 1600 coffee farmers who were interested in attending the monthly training modules. The study design for the evaluation then required that this pool of farmers be split into treatment and control groups, meaning those who would participate in the training, and those who wouldn’t (for now—they would be trained in later phases). The

trainings in Nyarubaka included 800 coffee farmers, randomly selected from the pool of 1600.

Randomization ensures that the treatment and comparison groups are equivalent at the beginning, mitigating concern for selection bias. But it cannot ensure that they remain comparable until the end of the program. Nor can it ensure that people comply with the treatment, or even the non-treatment, that they were assigned. Life also goes on after the randomization: other events besides the program happen between initial randomization and the end-line data collection. These events can reintroduce selection bias; they diminish the validity of the impact estimates and are threats to the integrity of the experiment. How can common threats to experimental integrity be managed?

EVALUATION DESIGN — THE EXPERIMENT AS PLANNED

As previously mentioned, the agronomy training evaluation consisted of 1600 farmers, half of which attended monthly training sessions, and the other half did not.

In addition, there was a census done of the entire sector to show us which households were coffee farmers and which ones were not. The census showed that there were 5400 households in Nyarubaka - 2400 non-coffee farming households and 3000 coffee farming households (1600 of which were already in our sample).

Each month a Technoserve farmer trainer would gather the farmers assigned to his/her group and conduct a training module on farming practices (e.g. weeding, pruning, bookkeeping, etc). The farmers were taught the best practices by using a practice plot so they could see and do exactly what the instructor was explaining.

To think about:

How can we be certain that the control group farmers did not attend the training too? What can be done to reduce this risk?

Since we have a census for Nyarubaka, how might this be helpful in at least controlling for or documenting any spillovers? (Think about what can be done at the trainings themselves).

What type of data might you need/want to try to control for any spillovers in this case?

What were other forms or opportunities for agronomy training in the area?

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 between 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?
5. In the Technoserve agronomy training example, why is it useful to randomly select from the farmers who signed up for the Technoserve training program, rather than amongst all the coffee farmers in the sector?

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 how much farmers adopt the recommendations and techniques from the agronomy trainings. Using a stylized example, let’s divide adoption of the techniques as follows:

Full adoption = score of 2

Partial adoption = score of 1

No adoption = score of 0

Let’s assume that there are 1800 farmers: 900 treatment farmers who receive the training and 900 comparison farmers who do not receive the training. After you randomize and collect some baseline data, you determine that the treatment and comparison groups are equivalent, meaning farmers from each of the three categories are equally represented in both groups.

Suppose protocol compliance is 100 percent: all farmers who are in the treatment go to the training and none of the farmers in the comparison attend the training. Let’s assume that all agronomy trainings end up with full adoption, scoring a 2. Let’s assume that there was a drought during this period, and those who adopted best-practices managed to protect their crops against damage. However, the farmers who have adoption level 0 see most of their crops perish, and members of the household enter the migrant labor market to generate additional income. The number of farmers in each treatment group, and each adoption category is shown for both the pre-adoption and post-adoption.

TABLE 1

Adoption Level	Pre-adoption		Post-adoption	
	T	C	T	C
0	300	300	0	Dropped out
1	300	300	0	300
2	300	300	900	300
Total farmers in sample	900	900	900	600

1. At program end, what is the average adoption for the treatment group?
2. At program end, what is the average adoption for the comparison group?
3. What is the difference?
4. Is this outcome difference an accurate estimate of the impact of the program? Why or why not?
5. If it is not accurate, does it overestimate or underestimate the impact?
6. How can we get a better estimate of the program’s impact?
7. Besides level of adoption, the Technoserve agronomy training evaluation also looked at outcome measures such as yields. Would differential attrition (i.e. differences in drop-outs between treatment and comparison groups) bias either of these outcomes? How?
8. Would the impacts on these final outcome measures be underestimated or overestimated?
9. You may know of other research designs to measure impact, such as the non-experimental or quasi-experimental methodologies (e.g. Pre-post, difference-in-difference, regression discontinuity, instrumental variables (IV), etc.) Is the threat of attrition unique to randomized evaluations?

**MANAGING PARTIAL COMPLIANCE—
WHEN THE TREATMENT DOES NOT
ACTUALLY GET TREATED OR THE
COMPARISON GETS TREATED**

Some people assigned to the treatment may in the end not actually get treated. In an after-school tutoring program, for example, some children assigned to receive tutoring may simply not show up for tutoring. And the others assigned to the comparison may obtain access to 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 that farmers who have adoption level 0 are too risk averse to adopt the techniques they learn at the training. Farmers believe that there is no way for them to adopt the techniques that are described in early trainings and stop attending. Consequently, none of the treatment farmers with adoption level 0 increased their adoption and remained at level 0 at the end of the program. No one assigned to comparison had attended the trainings. All the farmers in the sample at the beginning of the program were followed up.

TABLE 2

	Pre-adoption		Post-adoption	
	T	C	T	C
Adoption Level				
0	300	300	300	300
1	300	300	0	300
2	300	300	600	300
Total # farmer in the sample	900	900	900	900

1. Calculate the impact estimate based on the original group assignments.
 - a. Is this an unbiased measure of the effect of the program?
 - b. How might this estimate be useful when thinking about whether to expand the program to a larger number of its target audience?
2. You are interested in learning the effect of treatment on those actually treated (“treatment on the treated” (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 farmers who attended the training.

- a. Is this advice sound? Why or why not?
3. Another colleague says that it's not a good idea to drop the farmers who stopped attending the trainings entirely; you should use them but consider them as part of the control group.
 - 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?

1. How do you imagine spillovers might occur in agronomy training?
2. What types of mechanisms can you think of that could be used to reduce or manage spillovers?

Discussion Topic 5

Measuring spillovers

1. Can you think of ways to design the experiment explicitly to measure the spillovers of the agronomy training?

MANAGING SPILLOVERS—WHEN THE COMPARISON, ITSELF UNTREATED, BENEFITS FROM THE TREATMENT BEING TREATED

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

Discussion Topic 4

Managing spillovers

In the Technoserve agronomy training evaluation, randomization was at the farmer level, meaning that while one farmer might have been selected to be in the training, his neighbor didn't have the same fortunes during the randomization process.

Depending on the evaluation and the nature of the program, it might be more challenging to prevent spillovers of agronomic knowledge between friends, than it is for delivering hard tangible objects in farmers' hands, like a weighing scale or calendar to maintain harvest records.



EXERCISE A: UNDERSTANDING RANDOM SAMPLING AND THE LAW OF LARGE NUMBERS

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

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



EXERCISE B: HOW TO DO RANDOM ASSIGNMENT USING MS EXCEL

TABLE OF CONTENTS

Introduction.....	28
Part 1: Simple Randomization.....	28
Part 2: Complete randomization.....	32
Part 3: Stratified randomization.....	35

INTRODUCTION

Like most spreadsheet programs, MS Excel can generate random numbers on command. MS Excel has two native random-number-generating functions. The first, =RAND(), creates a *continuous* random number between 0 and 1—it could be any number of 9 decimal places between 0 and 1. The second, =RANDBETWEEN(*bottom*, *top*) creates *integers* between any two integer values within a range, where you specify the *bottom* and *top* of that range.

PART 1: SIMPLE RANDOMIZATION

Say we had a list of schools and we wanted to assign them to treatment or control based on a coin flip (heads = treatment and tails = control). We can do this by randomly generating the value of 0 or 1 using the RANDBETWEEN function, and choosing 0 and 1 as the range. We could then assign all schools with 0 to the control group, and all schools with 1 to the treatment group (or vice versa). This is equivalent to a coin flip where 0 represents tails and 1 represents control. Equivalently, we could produce a *continuous* random number for each observation and assign those with (say) random number greater than or equal to 0.5 to treatment and smaller than 0.5 to control.

The illustration on the following page shows how to do this step-by-step.

We have a list of all schools

	A	B	C	D	E
1	SchoolID	SchoolName	Pre-test score	Random #	T-C
2	101	Babajipura G.M.M.Kumar shala No. 1	34.12		
3	103	Babajipura Kanya Shala No. 3	25.49		
4	107	Babajipura Mishra Shala No. 7	12.60		
5	108	Babajipura Mishra Shala No. 8	20.77		
6	112	Babajipura Marathi Mishra Shala No. 12	37.69		
7	113	Babajipura Kanya Shala No. 13	32.71		
8	114	Babajipura Mishra Shala No. 14	16.32		
9	117	Babajipura Kumar Shala No. 17	20.12		
10	118	Babajipura Mishra Shala No. 18	28.05		
11	119	Babajipura Mishra Shala No. 19	21.29		
12	120	Babajipura Mishra Shala No. 20	26.34		
13	121	Babajipura Mishra Shala No. 21	16.36		
14	125	Babajipura Kumar Shala No. 25	21.32		
15	126	Babajipura Kanya Shala No. 26	25.25		
16	127	Babajipura Mishra Shala No. 27	26.43		
17	128	Babajipura Mishra Shala No. 28	29.38		
18	130	Babajipura Hindi Mishra Shala No. 30	18.21		
19	131	Babajipura Mishra Shala No. 31	20.70		
20	132	Babajipura Mishra Shala No. 32	34.72		
21	201	Fatehpura Kumar Shala No. 1	30.04		
22	202	Fatehpura Mishra Shala No. 2	19.53		
23	209	Fatehpura Mishra Shala No. 9	25.63		
24	210	Fatehpura Kanya Shala No. 10	18.96		
25	211	Fatehpura Mishra Shala No. 11	21.11		
26	213	Fatehpura Kumar Shala No. 13	18.09		
27	215	Fatehpura Hindi Mishra Shala No. 15	23.27		
28	216	Fatehpura Mishra Shala No. 16	22.74		
29	218	Fatehpura Mishra Shala No. 18	15.08		
30	219	Fatehpura Mishra Shala No. 19	25.37		
31	301	N. Sayajiganj Mishra Shala No. 1 (center)	18.27		
32	303	N. Sayajiganj Marathi Mishra Shala No. 3	31.90		
33	305	Sayajiganj Mishra Shala No. 5	19.00		
34	306	Sayajiganj Kumar Shala No. 6	20.81		
35	307	Sayajiganj Mishra Shala No. 7	47.18		

Step 1: Assign a random number to each school

The function RAND () is Excel's basic random number generator. To use it, go to Column D and type

=RAND()

in each cell, adjacent to each school name. Or you can type this function in the top row (row 2) and simply copy and paste to the entire column, or click and drag.

	A	B	C	D	E
1	SchoolID	SchoolName	Pre-test score	Random #	T-C
2	101	Babajipura G.M.M.Kumar shala No. 1	34.12	=RAND()	
3	103	Babajipura Kanya Shala No. 3	25.49		
4	107	Babajipura Mishra Shala No. 7	12.60		
5	108	Babajipura Mishra Shala No. 8	20.77		
6	112	Babajipura Marathi Mishra Shala No. 12	37.69		
7	113	Babajipura Kanya Shala No. 13	32.71		
8	114	Babajipura Mishra Shala No. 14	16.32		
9	117	Babajipura Kumar Shala No. 17	20.12		

Typing =RAND() puts a 9-digit random number between 0 and 1 in the cell.

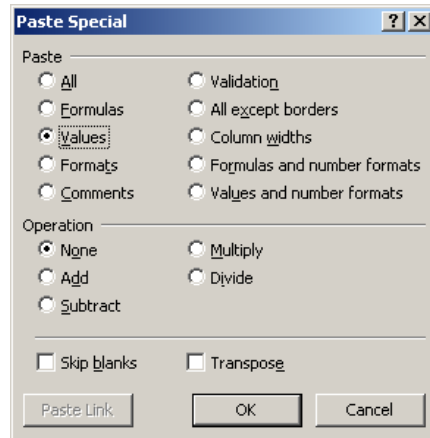
	A	B	C	D	E
1	SchoolID	SchoolName	Pre-test score	Random #	T-C
2	101	Babajipura G.M.M.Kumar shala No. 1	34.12	0.0789009	
3	103	Babajipura Kanya Shala No. 3	25.49	0.8999008	
4	107	Babajipura Mishra Shala No. 7	12.60	0.4359626	
5	108	Babajipura Mishra Shala No. 8	20.77	0.1408828	
6	112	Babajipura Marathi Mishra Shala No. 12	37.69	0.9634844	
7	113	Babajipura Kanya Shala No. 13	32.71	0.2145561	
8	114	Babajipura Mishra Shala No. 14	16.32	0.2558066	
9	117	Babajipura Kumar Shala No. 17	20.12	0.0169244	
10	118	Babajipura Mishra Shala No. 18	28.05	0.0655376	
11	119	Babajipura Mishra Shala No. 19	21.29	0.2724011	
12	120	Babajipura Mishra Shala No. 20	26.34	0.7489921	
13	121	Babajipura Mishra Shala No. 21	16.36	0.0268576	
14	125	Babajipura Kumar Shala No. 25	21.32	0.0661789	
15	126	Babajipura Kanya Shala No. 26	25.25	0.6946606	
16	127	Babajipura Mishra Shala No. 27	26.43	0.5000895	
17	128	Babajipura Mishra Shala No. 28	29.38	0.642025	
18	130	Babajipura Hindi Mishra Shala No. 30	18.21	0.8219122	
19	131	Babajipura Mishra Shala No. 31	20.70	0.7963628	
20	132	Babajipura Mishra Shala No. 32	34.72	0.5042257	
21	201	Fatehpura Kumar Shala No. 1	30.04	0.9492957	
22	202	Fatehpura Mishra Shala No. 2	19.53	0.9989293	
23	209	Fatehpura Mishra Shala No. 9	25.63	0.2719192	
24	210	Fatehpura Kanya Shala No. 10	18.96	0.5246963	
25	211	Fatehpura Mishra Shala No. 11	21.11	0.2142812	
26	213	Fatehpura Kumar Shala No. 13	18.09	0.6100928	
27	215	Fatehpura Hindi Mishra Shala No. 15	23.27	0.8909558	
28	216	Fatehpura Mishra Shala No. 16	22.74	0.2995547	
29	218	Fatehpura Mishra Shala No. 18	15.08	0.2206103	

Step 2: Copy the cells in Column D, 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 F9 key.)

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 D. Then right-click anywhere in the highlighted column, and choose "Copy". Then, right-click anywhere in that column and choose "Paste Special." The "Paste Special" window will appear. Click on "Values".



Step 3: Assign treatment/control status for each group

Now use the IF function to assign schools to treatment and control. Go to column E and type

`=IF(D2>=0.5,"T","C")`

And click and drag (or copy and paste) to the rest of the column. This will enter a "T" for schools that have a random number greater than or equal to 0.5 and "C" for schools with random number less than 0.5.

	A	B	C	D	E	F	G
1	SchoolID	SchoolName	Pre-test score	Random #	T-C		
2	215	Fatehpura Hindi Mishra Shala No. 15	23.27	0.4280308	=IF(D2>=0.5,"T","C")		
3	511	Shaher Vibhag Mishra Shala No. 11	22.37	0.4977787	C		
4	341	Sayajiganj Kanya Shala No. 41	24.57	0.5068542	T		
5	632	Wadi Kanya Shala No. 32	30.52	0.6071675	T		
6	514	Shaher Vibhag Mishra Shala No. 14	25.29	0.9297094	T		
7	626	Wadi Hindi Mishra Shala No. 26	41.10	0.3811165	C		
8	345	Sayajiganj Mishra Shala No. 45	20.33	0.0250151	C		
9	210	Fatehpura Kanya Shala No. 10	18.96	0.3442701	C		
10	622	Wadi Mishra Shala No. 22	21.90	0.0106587	C		
11	101	Babajipura G.M.M.Kumar shala No. 1	34.12	0.8055242	T		
12	315	Sayajiganj Hindi Mishra Shala No. 15	28.60	0.9751691	T		
13	313	Sayajiganj Mishra Shala No. 13	27.39	0.9867175	T		

Your list of schools has now been randomly assigned to treatment and control!

Is the number of schools in in both groups the same? We also have the average pre-test scores for each school. Does the average pre-test score look balanced between the two groups?

	Treatment	Control	Difference
Number of Schools	63	59	-4
Average Pre-test Score	25.62	26.92	1.30

Note, however, that the number of schools in treatment and control will vary each time you re-randomize, as will the average pre-test score. To check this, repeat step 1, but this time instead of copy pasting values, press the F9 key to re-randomize. Re-randomize 10 times and see what happens to the number of schools and the average pre-test score in each group.

Does the number of schools change when you re-randomize? Does the average pre-test score look balanced every time you re-randomize?

Try the above steps using the RANDBETWEEN() function instead of the RAND() function. Do you expect significantly different results? How does the “IF” function change?

PART 2: COMPLETE RANDOMIZATION

Say we had a list of schools and wanted to assign exactly half of them to treatment and half to control

Step 1: Assign a random number to each school

Go to Column D and type:

=RAND()

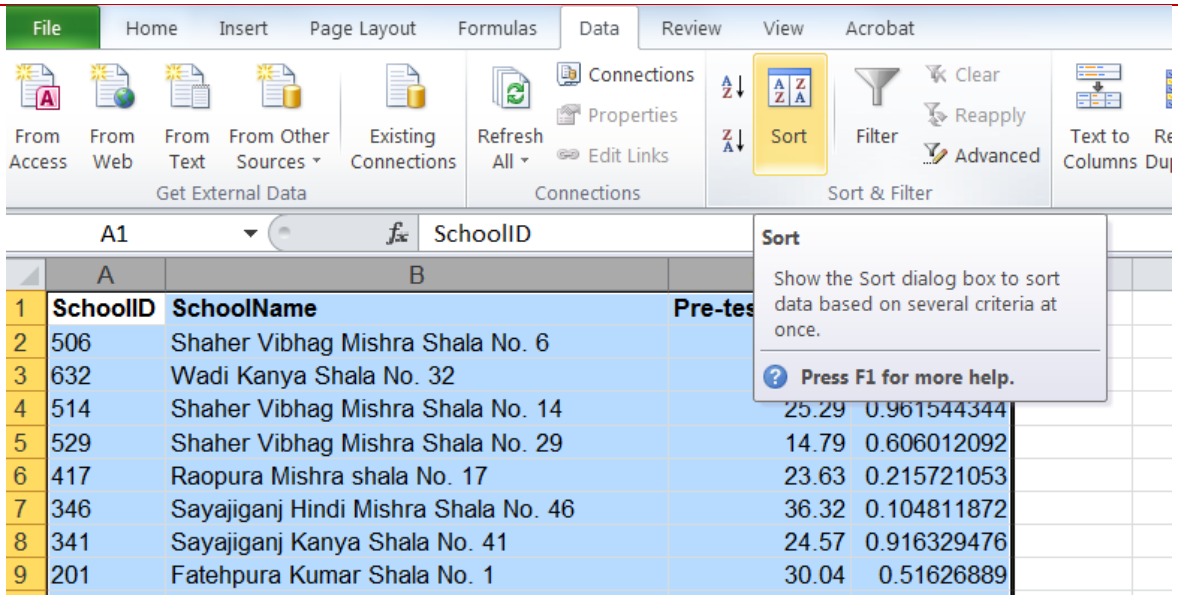
And click and drag (or copy and paste) to the entire column.

Step 2: Copy the cells in Column D, then paste the values over the same cells

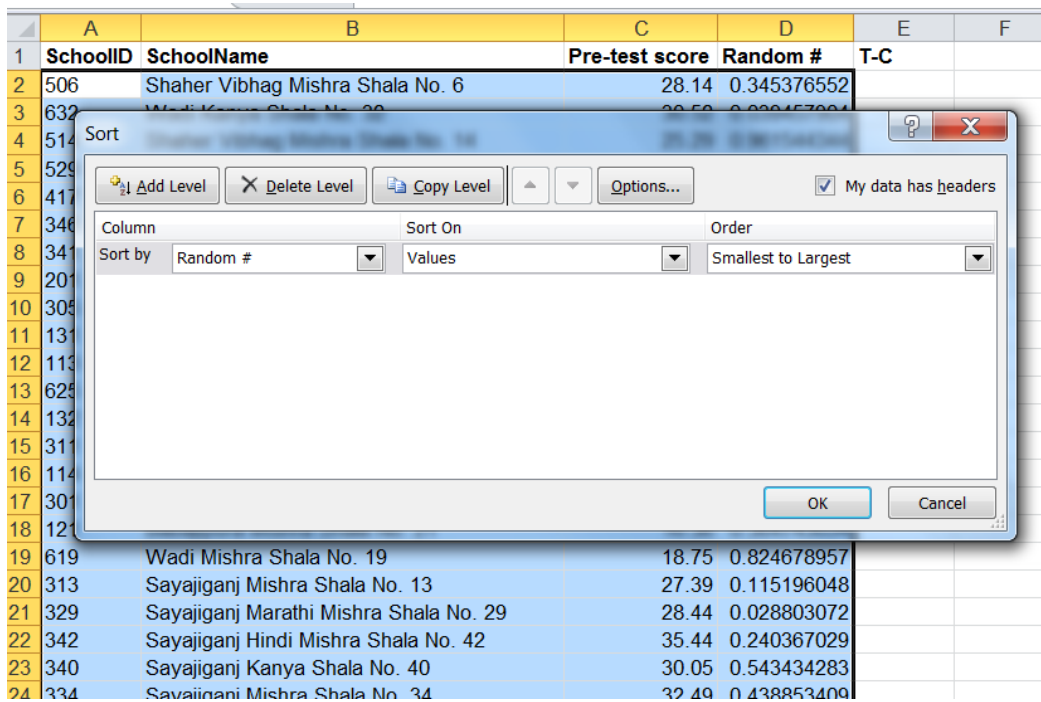
Highlight all values in Column D. Then right-click anywhere in the highlighted column, and choose “Copy”. Then, right-click anywhere in that column and choose “Paste Special.”

Step 3: Sort the columns in either descending or ascending order of Column D

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



A Sort box will pop up.



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

There! You have a randomly sorted list.

	A	B	C	D	E
1	SchoolID	SchoolName	Pre-test score	Random #	T-C
2	409	Raopura Marathi Mishra Shala No. 9	35.15	0.009954911	
3	323	Sayajiganj Mishra Shala No. 23	24.18	0.014434557	
4	510	Shaher Vibhag Mishra Shala No. 10	42.07	0.021591901	
5	529	Shaher Vibhag Mishra Shala No. 29	14.79	0.02215352	
6	611	Wadi Marathi Mishra Shala No. 11	36.31	0.027430569	
7	307	Sayajiganj Mishra Shala No. 7	47.18	0.041933781	
8	503	Shaher Vibhag Kanya Shala No. 3	22.15	0.043774186	
9	401	Raopura Mishra Shala No. 1	26.21	0.048703656	
10	326	Sayajiganj Hindi Mishra Shala No. 26	33.54	0.063108687	
11	411	Raopura Kumar Shala No. 11	20.62	0.06761041	
12	127	Babajipura Mishra Shala No. 27	26.43	0.073245278	
13	638	Wadi Hindi Mishra Shala No. 38	32.19	0.073990044	
14	119	Babajipura Mishra Shala No. 19	21.29	0.076187741	
15	132	Babajipura Mishra Shala No. 32	34.72	0.076497148	
16	322	Sayajiganj Mishra Shala No. 22	12.76	0.087452794	
17	128	Babajipura Mishra Shala No. 28	29.38	0.088030307	
18	618	Wadi Kumar Shala No. 18	23.77	0.093828679	
19	349	Sayajiganj Mishra Shala No. 49	40.16	0.10107943	
20	637	Wadi Mishra Shala No. 37	25.26	0.109629264	
21	311	Sayajiganj Mishra Shala No. 11	26.24	0.110998693	
22	501	Shaher Vibhag Mishra Shala No. 1	27.31	0.127617726	
23	639	Wadi Mishra Shala No. 39	20.49	0.127796284	
24	508	Shaher Vibhag Mishra Shala No. 8	22.19	0.134358453	
25	348	Sayajiganj Mishra Shala No. 48	28.23	0.140845208	
26	338	Sayajiganj Kanya Shala No. 38	29.22	0.144582844	
27	617	Wadi Marathi Mishra Shala No. 17	35.67	0.172061028	
28	213	Fatehpura Kumar Shala No. 13	18.09	0.17400346	
29	314	Sayajiganj Kumar Shala No. 14	22.93	0.182726341	
30	623	Wadi Kanya Shala No. 23	28.91	0.186591564	
31	624	Wadi Mishra Shala No. 24	21.52	0.208067391	
32	344	Sayajiganj Mishra Shala No. 44	25.91	0.209175575	
33	633	Wadi Kanya Shala No. 33	29.51	0.227463469	

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.”

Step 4: Assign treatment/control status for each group

There are two ways to do this. To do this manually, in column E, type “T” for the first half of the rows (rows 2–63) and for the second half of the rows (rows 62–123), type “C”. You can also do this by using the IF and MEDIAN functions. In Column E type:

```
=IF(D2<=MEDIAN($D$2:$D$123),"T","C")
```

And click and drag (or copy and paste) to the entire column. This will enter a “T” for schools that are below or at the median of the random number and a “C” for schools that are above it.

	A	B	C	D	E	F	G
1	SchoolID	SchoolName	Pre-test score	Random #	T-C		
2	409	Raopura Marathi Mishra Shala No. 9	35.15	0.009954911	=IF(ROW()<=62,"T","C")		
3	323	Sayajiganj Mishra Shala No. 23	24.18	0.014434557			
4	510	Shaher Vibhag Mishra Shala No. 10	42.07	0.021591901			
5	529	Shaher Vibhag Mishra Shala No. 29	14.79	0.02215352			
6	611	Wadi Marathi Mishra Shala No. 11	36.31	0.027430569			
7	307	Sayajiganj Mishra Shala No. 7	47.18	0.041933781			
8	503	Shaher Vibhag Kanya Shala No. 3	22.15	0.043774186			
9	401	Raopura Mishra Shala No. 1	26.21	0.048703656			
10	326	Sayajiganj Hindi Mishra Shala No. 26	33.54	0.063108687			
11	411	Raopura Kumar Shala No. 11	20.62	0.06761041			
12	127	Babajipura Mishra Shala No. 27	26.43	0.073245278			
13	638	Wadi Hindi Mishra Shala No. 38	32.19	0.073990044			
14	119	Babajipura Mishra Shala No. 19	21.29	0.076187741			
15	132	Babajipura Mishra Shala No. 32	34.72	0.076497148			
16	322	Sayajiganj Mishra Shala No. 22	12.76	0.087452794			
17	128	Babajipura Mishra Shala No. 28	29.38	0.088030307			
18	618	Wadi Kumar Shala No. 18	23.77	0.093828679			
19	349	Sayajiganj Mishra Shala No. 49	40.16	0.10107943			
20	637	Wadi Mishra Shala No. 37	25.26	0.109629264			
21	311	Sayajiganj Mishra Shala No. 11	26.24	0.110998693			
22	501	Shaher Vibhag Mishra Shala No. 1	27.31	0.127617726			
23	639	Wadi Mishra Shala No. 39	20.49	0.127796284			
24	508	Shaher Vibhag Mishra Shala No. 8	22.19	0.134358453			
25	348	Sayajiganj Mishra Shala No. 48	28.23	0.140845208			

Now select columns A through E and re-sort your list back in order of “SchoolID.” You’ll see that your schools have been randomly assigned to treatment and control groups.

Is the number of schools in both groups the same? Does the average pre-test score look balanced between the two groups?

Note that the number of schools in treatment and control will remain the same each time you re-randomize. This is because you are making sure that you always assign half of them to treatment and half to control. To check this, repeat step 1, but this time instead of copy pasting values, press the F9 key to re-randomize. Notice that the formula in column E will automatically recalculate the median each time and re-assign treatment and control status. Re-randomize 10 times and see what happens to the number of schools and the average pre-test score in each group.

Does the number of schools change when you re-randomize? Does the average pre-test score look balanced every time you re-randomize?

PART 3: STRATIFIED RANDOMIZATION

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

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.

Step 1: Assign a random number to each school

Go to Column F and type:

```
=RAND()
```

And click and drag (or copy and paste) to the entire column.

Step 2: Copy the cells in Column F, then paste the values over the same cells

Highlight all values in Column F. Then right-click anywhere in the highlighted column, and choose “Copy”. Then, right-click anywhere in that column and choose “Paste Special.”

Step 3: Sort by strata and then by random number

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

	A	B	C	D	E	F	G
	SchoolID	SchoolName	Pre-test score	Language	Gender	Random #	T-C
2	603	Wadi Mishra Shala No. 3	14.76	Gujarati	Co-ed	0.069825549	
3	625	Wadi Kumar Shala No. 25	28.90	Gujarati	Boys	0.32824482	
4	107						605
5	349						508
6	425						405
7	308						666
8	502						479
9	624						604
10	317						688
11	524						532
12	405						123
13	632						909
14	619						627
15	639						043
16	506						528
17	118						568
18	602						371
19	131	Dabjipura Mishra Shala No. 31	20.70	Gujarati	Co-ed	0.000000000	686
20	408	Raopura Kanya Shala No. 8	25.03	Gujarati	Girls	0.836292353	
21	316	Sayajiganj Kanya Shala No. 16	25.08	Gujarati	Girls	0.570278699	
22	417	Raopura Mishra shala No. 17	23.63	Gujarati	Co-ed	0.730099722	
23	601	Wadi Mishra Shala No. 1	36.44	Gujarati	Co-ed	0.085829273	
24	529	Shaheer Vibhag Mishra Shala No. 29	14.79	Gujarati	Co-ed	0.496626789	
25	637	Wadi Mishra Shala No. 37	25.26	Gujarati	Co-ed	0.903677425	
26	128	Bahainura Mishra Shala No. 28	29.38	Gujarati	Co-ed	0.285842087	

Step 4: Assign treatment/control status for each group

There are two ways to do this. To do this manually, in column G, within each languages category, type “T” for the first half of the rows, and “C” for the second half. You can also do this by using the IF and MEDIAN functions. In Column G type and instead of just hitting ENTER, hit CTRL+SHIFT+ENTER in order to tell Excel that it’s an array formula:

```
=IF(F2<MEDIAN(IF($D$2:$D$123=D2,$F$2:$F$123)),"T","C")
```

Click and drag (or copy and paste) to the entire column. This will enter a “T” for schools that are below or at the median of the random number and a “C” for schools that are above it for *each* language category.

Is the total number of schools in both groups the same? Is the number of schools for each language category for both groups the same? Does the average pre-test score look balanced between the two groups?

Note that the total number of schools and the number of schools for each language category in treatment and control will remain the same each time you re-randomize. To check this, repeat step 1, but this time instead of copy pasting values, press the F9 key to re-randomize. Notice that the formula in column E will automatically recalculate the median for each category every time and re-assign treatment and control status. Re-randomize 10 times and see what happens to the number of schools and the average pre-test score in each group.

Does the total number of schools change when you re-randomize? Does the number of schools for each language category change? Does the average pre-test score look balanced every time you re-randomize?



EXERCISE C: HOW TO DO POWER CALCULATIONS IN OPTIMAL DESIGN SOFTWARE

TABLE OF CONTENTS

Key Vocabulary	38
Introduction	39
Using the Optimal Design Software	39
Estimating Sample Size for a Simple Experiment.....	44
Some Wrinkles: Limited Resources and Imperfect Compliance.....	48
Clustered Designs.....	49

KEY VOCABULARY

1. POWER: The likelihood that, when a program/treatment has an effect, you will be able to distinguish the effect from zero i.e. from a situation where the program has no effect, 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 certain outcome indicators of interest (for instance, test scores in an education program.)

3. STANDARD DEVIATION: For a particular indicator, a measure of the variation (or spread) of a sample or population. Mathematically, this is the square root of the variance.

4. STANDARDIZED EFFECT SIZE: A standardized (or normalized) measure of the [expected] magnitude of the effect of a program. Mathematically, it is the difference between the treatment and control group (or between any two treatment arms) for a particular outcome, divided by the standard deviation of that outcome in the control (or comparison) group.

5. CLUSTER: The unit of observation at which a sample size is randomized (e.g. school), each of which typically contains several units of observation that are measured (e.g. students). Generally, observations that are highly correlated with each other should be clustered and the estimated sample size required should be measured with an adjustment for clustering.

6. INTRA-CLUSTER CORRELATION COEFFICIENT (ICC): A measure of the correlation between observations within a cluster. For instance, if your experiment is clustered at the school level, the ICC would be the level of correlation in test scores for children in a given school relative to the overall correlation of students in all schools.

INTRODUCTION

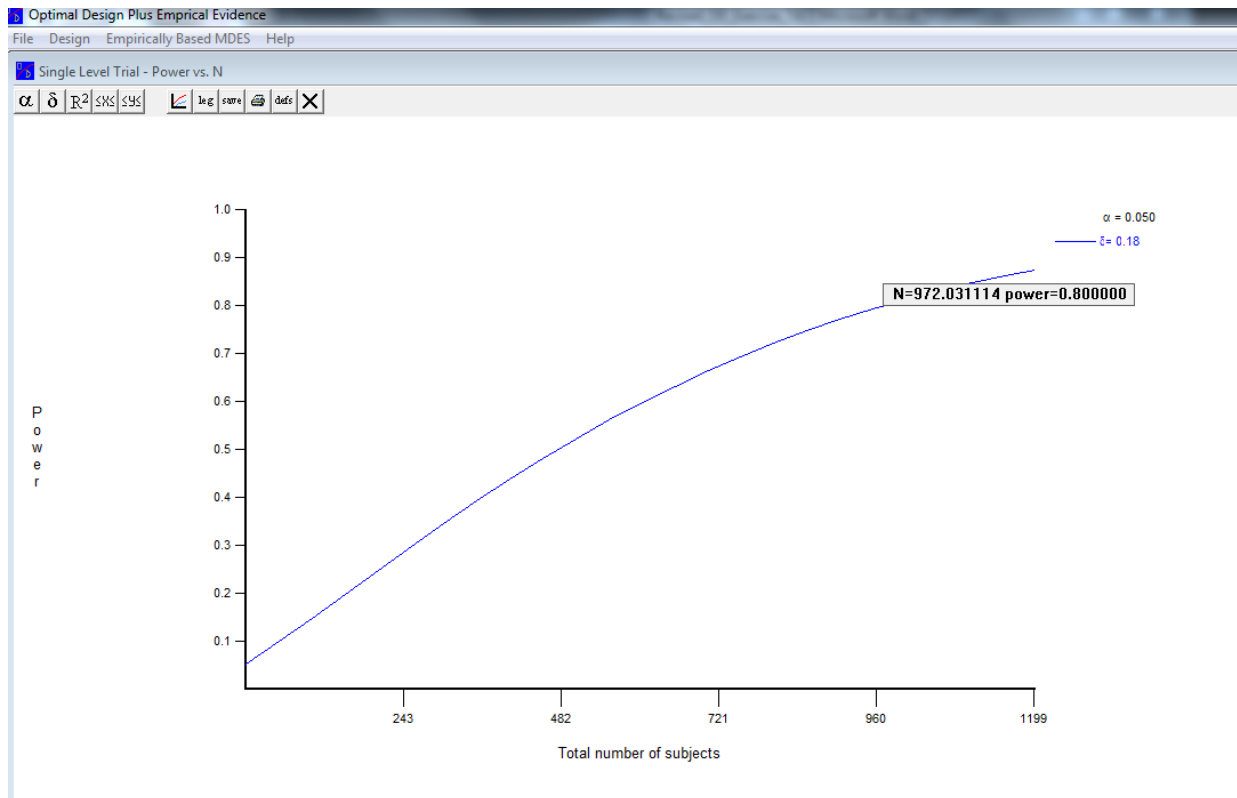
This exercise will help explain the trade-offs to power when designing a randomized trial. Should we sample every student in just a few schools? Should we sample a few students from many schools? How do we decide?

We will work through these questions by determining the sample size that allows us to detect a specific effect with at least 80 percent power, which is a commonly accepted level of power. Remember that power is the likelihood that when a program/treatment has an effect, you will be able to distinguish it from zero in your sample. Therefore at 80% power, if an intervention's impact is statistically significant at exactly the 5% level, then for a given sample size, we are 80% likely to detect an impact (i.e. we will be able to reject the null hypothesis.)

In going through this exercise, we will use the example of an education intervention that seeks to raise test scores. This exercise will demonstrate how the power of our sample changes with the number of school children, the number of children in each classroom, the expected magnitude of the change in test scores, and the extent to which children within a classroom behave more similarly than children across classrooms. We will use a software program called *Optimal Design*, developed by Stephen Raudenbush et al. with funding from the William T. Grant Foundation. Additional resources on research designs can be found on their web site.

USING THE OPTIMAL DESIGN SOFTWARE

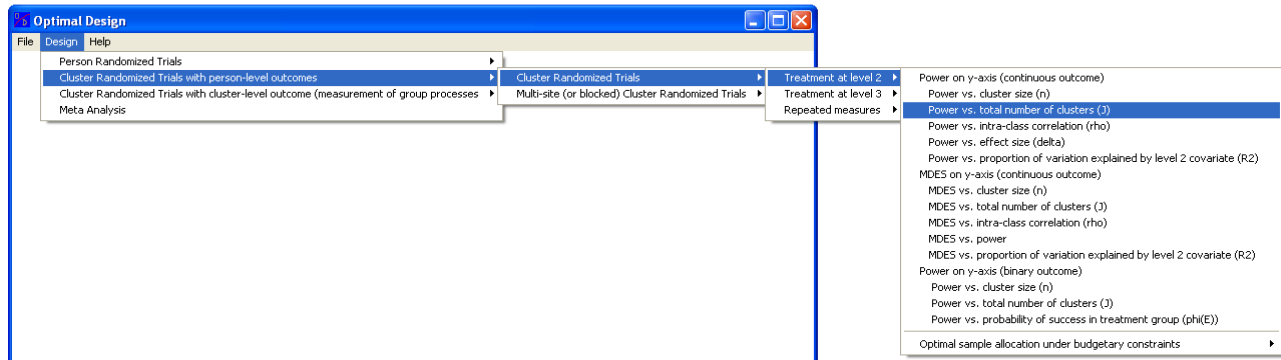
Optimal Design produces a graph that can show a number of comparisons: Power versus sample size (for a given effect), effect size versus sample size (for a given desired power), with many other options. The chart below shows power on the y-axis and sample size on the x-axis. In this case, we inputted an effect size of 0.18 standard deviations (explained in the example that follows) and we see that we need a sample size of 972 to obtain a power of 80%.



We will now go through a short example demonstrating how the OD software can be used to perform power calculations. If you haven't downloaded a copy of the OD software yet, you can do so from the following website (where a software manual is also available):

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

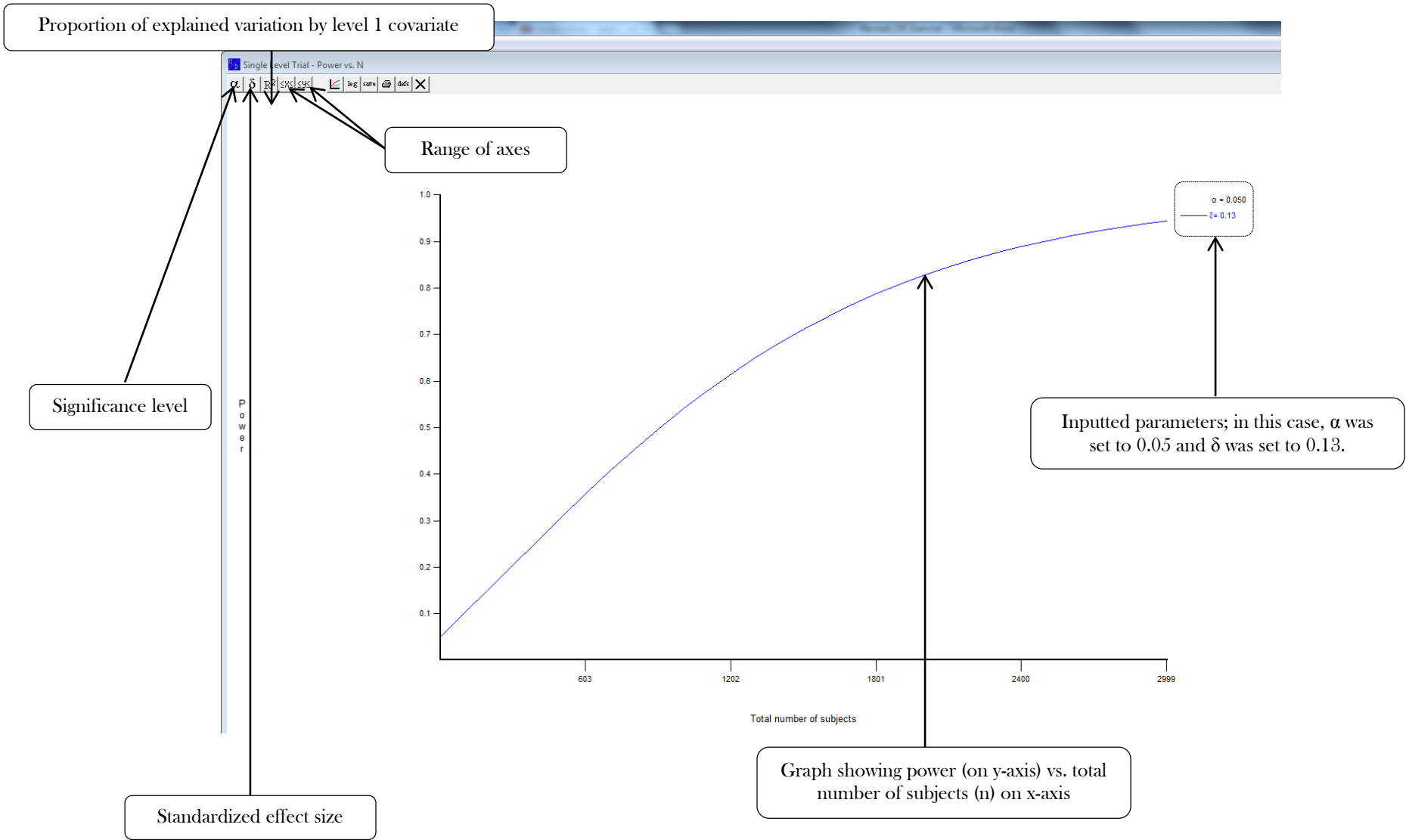
Running the HLM software file “od” should give you a screen which looks like the one below:



The various menu options under “Design” allow you to perform power calculations for randomized trials of various designs.

Let's work through an example that demonstrates how the sample size for a simple experiment can be calculated using OD. Follow the instructions along as you replicate the power calculations presented in this example, in OD. On the next page we have shown a sample OD graph, highlighting the various components that go into power calculations. These are:

- Significance level (α): For the significance level, typically denoted by α , the default value of 0.05 (i.e. a significance level of 95%) is commonly accepted.
- Standardized effect size (δ): Optimal Design (OD) requires that you input the standardized effect size, which is the effect size expressed in terms of a normal distribution with mean 0 and standard deviation 1. This will be explained in further detail below. The default value for δ is set to 0.200 in OD.
- Proportion of explained variation by level 1 covariate (R^2): This is the proportion of variation that you expect to be able to control for by including covariates (i.e. other explanatory variables other than the treatment) in your design or your specification. The default value for R^2 is set to 0 in OD.
- Range of axes (SxS and SyS): Changing the values here allows you to view a larger range in the resulting graph, which you will use to determine power.



We will walk through each of these parameters below and the steps involved in doing a power calculation. Prior to that though, it is worth taking a step back to consider what one might call the “paradox of power”. Put simply, in order to perfectly calculate the sample size that your study will need, it is necessary to know a number of things: the effect of the program, the mean and standard deviation of your outcome indicator of interest for the control group, and a whole host of other factors that we deal with further on in the exercise. However, we cannot know or observe these final outcomes until we actually conduct the experiment! We are thus left with the following paradox: In order to conduct the experiment, we need to decide on a sample size...a decision that is contingent upon a number of outcomes that we cannot know without conducting the experiment in the first place.

It is in this regard that power calculations involve making careful assumptions about what the final outcomes are likely to be – for instance, what effect you realistically expect your program to have, or what you anticipate the average outcome for the control group being. These assumptions are often informed by real data: from previous studies of similar programs, pilot studies in your population of interest, etc. The main thing to note here is that to a certain extent, power calculations are more of an art than a science. However, making wrong assumptions will not affect accuracy (i.e, will not bias the results). It simply affects the precision with which you will be able to estimate your impact. Either way, it is useful to justify your assumptions, which requires carefully thinking through the details of your program and context.

With that said, let us work through the steps for a power calculation using an example. Say your research team is interested in looking at the impact of providing students a tutor. These tutors work with children in grades 2, 3 and 4 who are identified as falling behind their peers. Through a pilot survey, we know that the average test scores of students before receiving tutoring is 26 out of 100, with a standard deviation of 20. We are interested in evaluating whether tutoring can cause a 10 percent increase in test scores.

1) Let’s find out the minimum sample that you will need in order to be able to detect whether the tutoring program causes a 10 percent increase in test scores. Assume that you are randomizing at the school level i.e. there are treatment schools and control schools.

I. What will be the mean test score of members of the control group? What will the standard deviation be?

Answer:

II. If the intervention is supposed to increase test scores by 10%, what should you expect the mean and standard deviation of the treatment group to be after the intervention? Remember, in this case we are considering a 10% increase in scores over the scores of the control group, which we calculated in part I.

Answer:

III. Optimal Design (OD) requires that you input the standardized effect size, which is the effect size expressed in terms of a normal distribution with mean 0 and standard deviation 1. Two of the most important ingredients in determining power are the *effect size* and the *variance* (or standard deviation). The standardized effect size basically combines these two ingredients into one number. The standardized effect size is typically denoted using the symbol δ (delta), and can be calculated using the following formula:

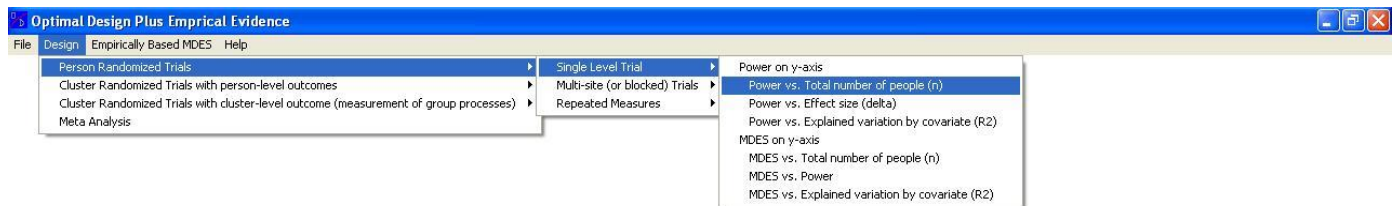
$$\delta = \frac{(\text{Treatment Mean} - \text{Control Mean})}{(\text{Standard Deviation})}$$

Using this formula, what is δ ?

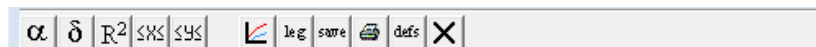
Answer:

IV. Now use OD to calculate the sample size that you need in order to detect a 10% increase in test scores. You can do this by navigating in OD as follows:

Design → Person Randomized Trials → Single Level Trial → Power vs. Total number of people (n)



There are various parameters that you will be asked to fill in:



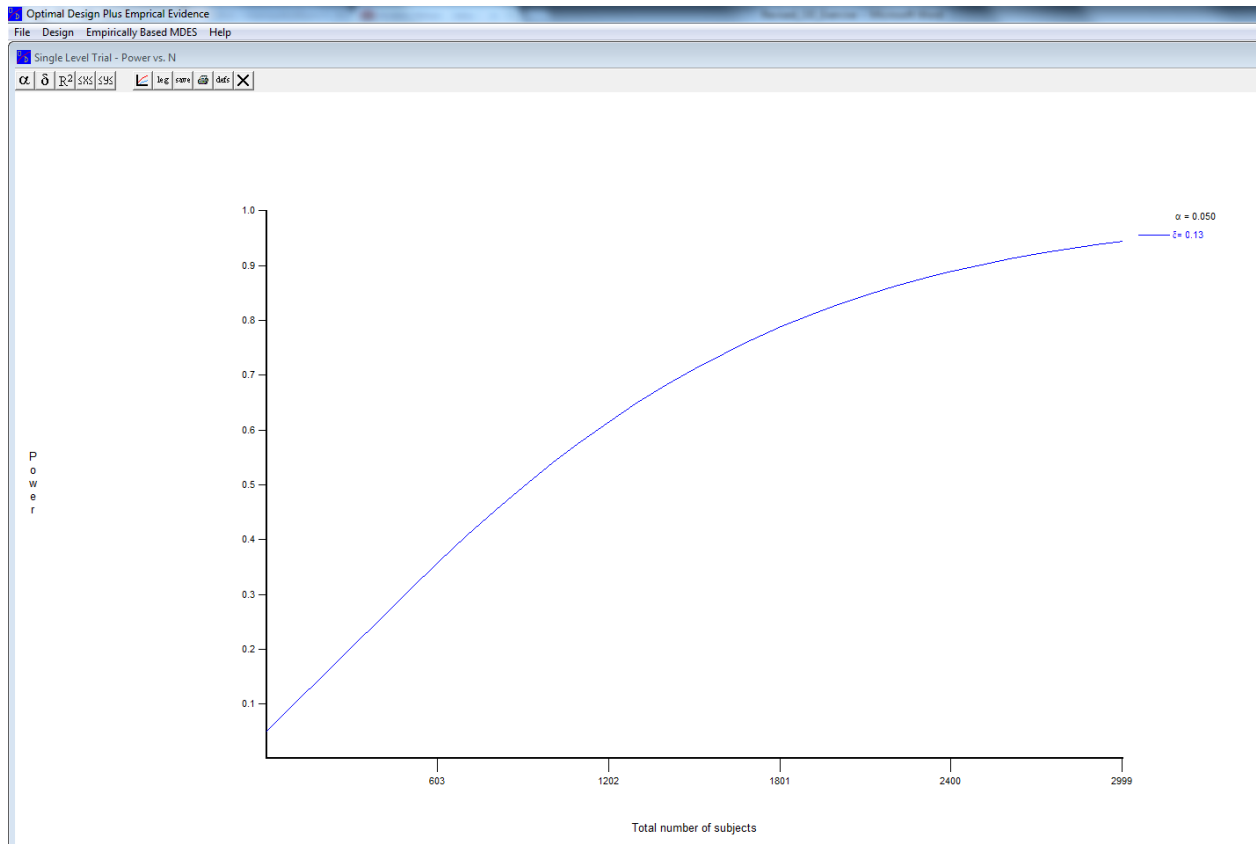
You can do this by clicking on the button with the symbol of the parameter. To reiterate, the parameters are:

- Significance level (α): For the significance level, typically denoted by α , the default value of 0.05 (i.e. a significance level of 95%) is commonly accepted.
- Standardized effect size (δ): The default value for δ is set to 0.200 in OD. However, you will want to change this to the value that we computed for δ in part C.
- Proportion of explained variation by level 1 covariate (R^2): This is the proportion of variation that you expect to be able to control for by including covariates (i.e. other explanatory variables other than the treatment) in your design or your specification. We will leave this at the default value of 0 for now and return to it later on.

- Range of axes (Δx and Δy): Changing the values here allows you to view a larger range in the resulting graph, which you will use to determine power; we will return to this later, but can leave them at the default values for now.

What will your total sample size need to be in order to detect a 10% increase in test scores at 80% power?

Answer:



ESTIMATING SAMPLE SIZE FOR A SIMPLE EXPERIMENT

All right, now it is your turn! For the parts A – I below, leave the value of R^2 at the default of 0 whenever you use OD; we will experiment with changes in the R^2 value a little later.

You decide that you would like your study to be powered to measure an increase in test scores of 20% rather than 10%. Try going through the steps that we went through in the example above. Let's find out the minimum sample you will need in order to detect whether the tutoring program can increase test scores by 20%.

A. What is the mean test score for the control group? What is the standard deviation?

Mean:

Standard deviation:

- B.** If the intervention is supposed to increase test scores by 20%, what should you expect the mean and standard deviation of the treatment group to be after the intervention?

Mean:

Standard deviation:

- C.** What is the desired standardized effect size δ ? Remember, the formula for calculating δ is:

$$\delta = \frac{(\text{Treatment Mean} - \text{Control Mean})}{(\text{Standard Deviation})}$$

δ :

- D.** Now use OD to calculate the sample size that you need in order to detect a 20% increase in test scores.

Sample size (n):

Treatment:

Control:

- E.** Is the *minimum* sample size required to detect a 10% increase in test scores larger or smaller than the minimum sample size required to detect a 20% increase in test scores? Intuitively, will you need larger or smaller samples to measure smaller effect sizes?

Answer:

- F.** Your research team has been thrown into a state of confusion! While one prior study led you to believe that a 20% increase in test scores is possible, a recently published study suggests that a more conservative 10% increase is more plausible. What sample size should you pick for your study?

Answer:

- G.** Both the studies mentioned in part F found that although average test scores increased after the tutoring intervention, the standard deviation of test scores also increased i.e. there was a larger spread of test scores across the treatment groups. To account for this, you posit that instead of 20, the standard deviation of test scores may now be 25 after the tutoring program. Calculate the new δ for an increase of 10% in test scores.

δ :

- H.** For an effect of 10% on test scores, does the corresponding standardized effect size increase, decrease, or remain the same if the standard deviation is 25 versus 20? Without plugging the values into OD, all other things being equal, what impact does a higher standard deviation of your outcome of interest have on your required sample size?

Answer:

- I.** Having gone through the intuition, now use OD to calculate the sample size required in order to detect a 10% increase in test scores, if the pre-intervention mean test scores are 26, with a standard deviation of 25.

Sample size (n):

Treatment:

Control:

- J.** One way by which you can increase your power is to include covariates i.e. control variables that you expect will explain some part of the variation in your outcome of interest. For instance, baseline, pre-intervention test scores may be a strong predictor of a child's post-intervention test scores; including baseline test scores in your eventual regression specification would help you to isolate the variation in test scores attributable to the tutoring intervention more precisely. You can account for the presence of covariates in your power calculations using the R^2 parameter, in which you specify what proportion of the eventual variation in your outcome of interest is attributable to your treatment condition.

Say that you have access to the pre-intervention test scores of children in your sample for the tutoring study. Moreover, you expect that pre-intervention test scores explain 50% of the variation in post-intervention scores. What size sample will you require in order to measure an increase in test scores of 10%, assuming standard deviation in test scores of 25, with a pre-intervention mean of 26. Is this more or less than the sample size that you calculated in part I?

Sample size (n):

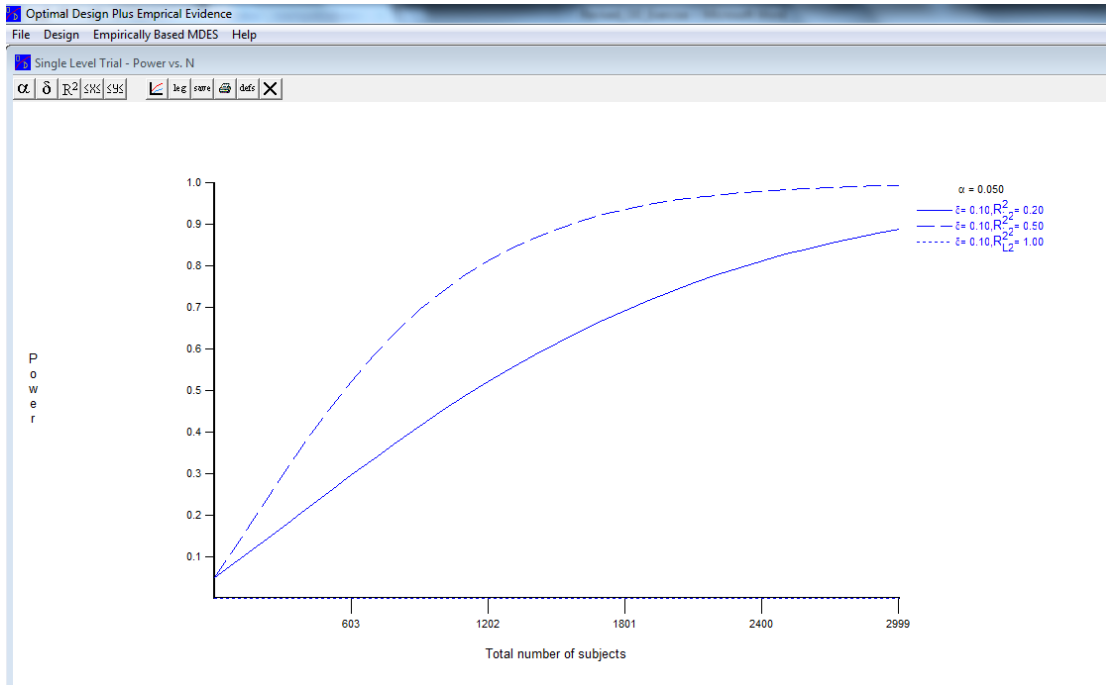
Treatment:

Control:

- K.** One of your colleagues on the research team thinks that 50% may be too ambitious an estimate of how much of the variation in test scores post-intervention is attributable to baseline scores. She suggests that 20% may be a better estimate. What happens to your required sample size when you run the calculations from part J with an R^2 of 0.200 instead of 0.500? What happens if you set R^2 to be 1.000?

Tip: You can enter up to 3 separate values on the same graph for the R^2 in OD; if you do, you will end up with a figure like the one on the following page:

Answer:



SOME WRINKLES: LIMITED RESOURCES AND IMPERFECT COMPLIANCE

- L.** You find out that you only have enough funds to survey 1,200 children. Assume that you do not have data on baseline covariates, but know that pre-intervention test scores were 26 on average, with a standard deviation of 20. What standardized effect size (δ) would you need to observe in order to survey a maximum of 1,200 children and still retain 80% power? Assume that the R^2 is 0 for this exercise since you have no baseline covariate data.

Hint: You will need to plot “Power vs. Effect size (delta)” in OD, setting “N” to 1,200. You can do this by navigating in OD as follows: Design → Person Randomized Trials → Single Level Trial → Power vs. Effect Size (delta). Then, click on the point of your graph that roughly corresponds to power = 0.80 on the y-axis.

$\delta =$

- M.** Your research team estimates that you will not realistically see more than a 10% increase in test scores due to the intervention. Given this information, is it worth carrying out the study on just 1,200 children if you are adamant about still being powered at 80%?

Answer:

- N.** Your research team is hit with a crisis: You are told that you cannot force people to use the tutors! After some small focus groups, you estimate that only 40% of schoolchildren would be interested in the tutoring services. You realize that this intervention would only work for a very limited number of schoolchildren. You do not know in advance whether students are likely to take up the tutoring service or not. How does this affect your power calculations?

Answer:

- O.** You have to “adjust” the effect size you want to detect by the proportion of individuals that actually gets treated. Based on this, what will be your “adjusted” effect size and the adjusted standardized effect size (δ) if you originally wanted to measure a 10% increase in test scores? Assume that your pre-intervention mean test score is 26, with a standard deviation of 20, you do not have any data on covariates, and that you can survey as many children as you want.

Hint: Keep in mind that we are calculating the average treatment effect for the entire group here. Thus, the lower the number of children that actually receives the tutoring intervention, the lower will be the measured effect size.

Answer:

- P. What sample size will you need in order to measure the effect size that you calculated in part O with 80% power? Is this sample bigger or smaller than the sample required when you assume that 100% of children take up the tutoring intervention (as we did in the example at the start)?

Sample size (n):

Treatment:

Control:

CLUSTERED DESIGNS

Thus far we have considered a simple design where we randomize at the *individual-level* i.e. school children are either assigned to the treatment (tutoring) or control (no tutoring) condition. However, spillovers could be a major concern with such a design: If treatment and control students are in the same *school*, let alone the same classroom, students receiving tutoring may affect the outcomes for students not receiving tutoring (through peer learning effects) and vice versa. This would lead us to get a biased estimate of the impact of the tutoring program.

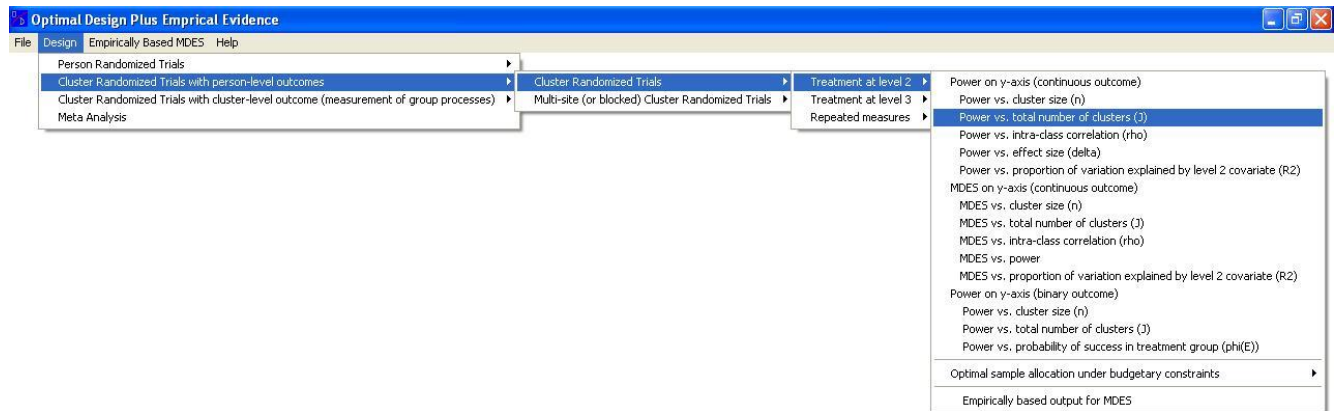
In order to preclude this, your research team decides that it would like to run a cluster randomized trial, randomizing at the *school-level* instead of the individual-level. In this case, each school forms a “cluster”, with all the students in a given school assigned to either the treatment condition, or the control one. Under such a design, the only spillovers that may show up would be across schools, a far less likely possibility than spillovers within schools.

Since the behavior of individuals in a given cluster will be correlated, we need to take an **intra-cluster or intra-class correlation (denoted by the Greek symbol ρ)** into account for each outcome variable of interest. Remember, ρ is a measure of the correlation between children within a given school (see key vocabulary at the start of this exercise.) 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 were in fact independent and there was zero difference between schools or any other factor that affected those students, then ρ would equal 0.

The ρ or ICC of a given variable is typically determined by looking at pilot or baseline data for your population of interest. Should you not have the data, another way of estimating the ρ is to look at other studies examining similar outcomes amongst similar populations. Given the inherent uncertainty with this, it is useful to consider a range of ρ s when conducting your power calculations (a sensitivity analysis) to see how sensitive they are to changes in ρ . We will look at this a little further on. While the ρ can vary widely depending on what you are looking at, values of less than 0.05 are typically considered low, values between 0.05-0.20 are considered to be of moderate size, and values above 0.20 are considered fairly high. Again, what counts as a low ρ and what counts as a high ρ can vary dramatically by context and outcome of interest, but these ranges can serve as initial rules of thumb.

Based on a pilot study and earlier tutoring interventions, your research team has determined that the ρ is 0.17. You need to calculate the total sample size to measure a 15% increase in test scores (assuming that test scores at the baseline are 26 on average, with a standard deviation of 20, setting R^2 to 0 for now). You can do this by navigating in OD as follows:

Design → Cluster Randomized Trials with person-level outcomes → Cluster Randomized Trials → Treatment at Level 2 → Power vs. total number of clusters (J)



In the bar at the top, you will see the same parameters as before, with an additional option for the intra-cluster correlation. Note that OD uses “n” to denote the cluster size here, not the total sample size. OD assigns two default values for the effect size (δ) and the intra-cluster correlation (ρ), so do not be alarmed if you see four lines on the chart. Simply delete the default values and replace them with the values for the effect size and intra-cluster correlation that you are using.

Q. What is the effect size (δ) that you want to detect here? Remember that the formula for calculating δ is:

$$\delta = \frac{(\text{Treatment Mean} - \text{Control Mean})}{(\text{Standard Deviation})}$$

δ :

R. Assuming there are 40 children per school, how many schools would you need in your clustered randomized trial?

Answer:

S. Given your answer above, what will the total size of your sample be?

Sample size:

Treatment:

Control:

T. What would the number of schools and total sample size be if you assumed that 20 children from each school were part of the sample? What about if 100 children from each school were part of the sample?

	20 children per school	40 children per school	100 children per school
Number of schools:		160	
Total no. of students:		6,400	

- U. As the number of clusters increases, does the total number of students required for your study increase or decrease? Why do you suspect this is the case? What happens as the number of children per school increases?

Answer:

- V. You realize that you had read the pilot data wrong: It turns out that the ρ is actually 0.07 and not 0.17. Now what would the number of schools and total sample size be if you assumed that 20 children from each school were part of the sample? What about if 40 or 100 children from each school were part of the sample?

	20 children per school	40 children per school	100 children per school
Number of schools:			
Total no. of students:			

- W. How does the total sample size change as you increase the number of individuals per cluster in part V? How do your answers here compare to your answers in part T?

Answer:

- X. Given a choice between offering the tutors to more children in each school (i.e. adding more individuals to the cluster) versus offering tutors in more schools (i.e. adding more clusters), which option is best *purely from the perspective of improving statistical power*? Can you imagine a situation when there will not be much difference between the two from the perspective of power?

Answer:

Guidelines for the Group Presentations

Learning in groups, facilitated by a Teaching Assistant (TA), is a central component of the course. Included in this course is a group project, where the participants, facilitated by their TA, will design a proposal for how they could evaluate a social programme of their own choosing, using a Randomised Evaluation.

The goal is for participants to plan a Randomised Evaluation that is both rigorous and pragmatic and, in doing so, consolidate and apply the knowledge learnt in the lectures. Ideally, the group presentation will be developed to the extent that it could be considered as the starting point for a real evaluation.

We encourage participants to choose a group project that is related to their work, even to the extent of it being a Randomised Evaluation that they would be interested in pursuing after the course, making the valuable advice from the Teaching Assistants and the J-PAL staff at the course further reaching and of greater benefit to the participants.


On the next page is a Power Point template that highlights the different steps that the proposal produced by the group should include. The template may be used as a guideline by the groups when they are preparing their presentation. The steps outlined in the Power Point template are:

1. Identifying and deciding on an intervention
2. Building the theory of change
3. Choosing the randomisation method
4. Power calculations
5. Identifying potential threats and solutions
6. Dissemination of results

Each group will present on the final day of the course to the presenters and participants of the course. Presentations should be kept to 15 minutes, allowing for 15 minutes discussion led by J-PAL affiliates and staff. We will provide groups with template slides for their presentation (see next page).

Group Presentation Template


Group work session 1

 J-PAL

Title


List your Team Members

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

 J-PAL


Background

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

 J-PAL


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.)

 J-PAL

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?

 J-PAL

Evaluation Design

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


Group work session 2

 J-PAL

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.

Group work session 3

 J-PAL

Potential challenges

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

 J-PAL

Results

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

Help Reading Regression Tables

Regression analysis is a statistical tool used to investigate relationships between variables. Usually, the investigator seeks to ascertain the effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate.

The table below is hypothetical, but draws on the example covered in case study 2, *Read India*. The intervention in this example is: teachers pulling children out of class for two hours a day to focus on basic numeracy and literacy skills. Researchers care about whether the intervention affects learning outcomes. Learning outcomes in this case are measured by test scores. Researchers hypothesize that learning outcomes depend on this intervention, thus “level” (test score) or “improvement” (the difference in test scores before and after the programme) are our **dependent variables**. The dependent variable is the variable we want to explain. Traditionally the dependent variable is placed horizontally in the first row. This is the case in the table below. The researcher wants to know how “reading classes” affect change in test score and test score level. “Reading Classes” is an **independent variable** -it explains learning outcomes. Independent variables are placed vertically in the first column. Usually researchers include other independent variables, variables that are likely to influence test scores, for example previous reading level, age, sex the type of school you went to or your parent’s literacy level. These **control variables** are included in order to describe test scores as precisely as possible, which enables us the model to create a more accurate relationship between test scores and reading classes.

Correlation coefficients are found in the cells in the inside of the table which are not in brackets. We analyse each correlation coefficient by looking at which row and column it is in. For example 0.68 is in the “Reading Class” row and the “Level” column. This means the number describes the relationship between “Reading Class” and “Level”. There are two things we look at when we analyse the number **magnitude** and **sign**.

- **MAGNITUDE:** A large number indicates a strong relationship between the independent variable (reading classes) and the dependent variable (test scores).
- **SIGN:** The sign indicates whether there is a negative or positive relationship between independent and dependent variable. For example a positive number indicates there is a positive relationship between reading classes and test scores, attending reading classes is associated with an increase in test scores, while a negative sign indicates that classes are associated with a decrease in test scores.

Standards errors are found in brackets beneath the correlation coefficients. The standard errors indicate how well the model fits the data, which is how well the independent variables explain the dependent variables. Small standard errors suggest that the model describes the dependent variable well.

Statistical significance is indicated using stars. One star indicates that the variable is statistically significant at the 10% level. This means that we are 90% certain that the coefficient represents the correct relationship between the dependent and independent variable, but there is a 10% chance that it could be spurious.

The table below points to the important parts of a regression table.

		Level		Improvement	
		(1)	(2)	(3)	(4)
Outcome variables: the primary outcomes we want to measure	Reading Classes	-0.68** (0.0829)	0.04 (0.1031)	0.24** (0.0628)	0.11 (0.1081)
	Previous Reading Level		0.71** (0.0215)		
Explanatory variables: factors that impact the outcome { Key explanatory variable: the program that is being evaluated { Control variables: all other variables that will affect the outcome variable	Age		0.00 (0.0182)		-0.01 (0.0194)
	Sex		-0.01 (0.0469)		0.05 (0.0514)
	Standard		0.02 (0.0174)		-0.08** (0.0171)
	Parents Literate		0.04 (0.0457)		0.13** (0.0506)
Coefficients: The average size of the impact that an explanatory variable has on an outcome variable.	Constant	2.82 (0.0239)	0.36 (0.2648)	0.37 (0.0157)	.75 (0.3293)
Statistical significance: statistical significance is usually indicated with stars in regression tables. 3 stars indicates highly significant results. 1 star indicates only slightly significant results. If there are no stars we cannot be confident that the variables has an impact.	School-type controls	No	Yes	No	Yes

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.

Welcome to Cape Town! Here are some Practical Tips

Taxis and Transport from Airport

There are metered taxis available to/from the airport. Standard rates are between 10 and 12 Rand per km. A trip from the airport to UCT / Rondebosch should not cost more than R200. Mention that you need a receipt before entering a cab.

Taxi services include:

Excite Cabs: 021 418 4444

Cabs on Call: 021 522 6103

Cab Xpress: 021 448 1616

Restaurants

Cape Town is known for its diverse array of dining and cuisine. Here is but a small list of well-known restaurants that you may wish to try.

Budget (Main meal under R60)

1) Eastern Food Bazaar

Cuisine: Indian, Chinese

Location: City Bowl

Contact: 021 461 2458

2) Mzolis

Cuisine: African, BBQ

Location: Gugulethu

Contact: 021 638 1355

3) Food Lovers Market

Cuisine: Deli, Buffet – Basically everything

Location: Claremont

Contact: 021 674 7836

4) Cocoa Wah Wah/ Cocoa Cha chi/Cocoa Oolah (Friday night after 7pm

Pasta Special- Pasta's for R39)

Look on website to find closest branch:

<http://www.cocoa.co.za/index.html>

5) Knead (R50 special)

All meals R50 from Monday to Friday between

4pm and 9pm

Look on website to find closest branch:

<http://kneadbakery.co.za/kneadcafessouthafrica.html#by>.

Medium price range
(Main meal between R60 and R100)

- 1) *Col Cacchio Pizzeria**
Cuisine: Pizza
Location: Claremont (Cavendish),
Camps Bay
Contact: 021 674 6387/ 021 438 2171

- 2) *Kirstenbosch Tea Room**
Cuisine: Coffee Shop
Location: Kristenbosch National
Botanical Gardens Newlands (Not for
dinner)
Contact: 021 797 4083

- 3) *Rhodes Memorial Restaurant**
Cuisine: Bistro, Coffee Shop
Location: Rhodes Memorial
Restaurant (Not for dinner)
Contact: 021 687 0000

- 5) *Fadela Williams**
Cuisine: Cape Malay
Location: Claremont
Contact: 021 671 0037

- 6) *Hussar Grill**
Cuisine: Grills
Location: Rondebosch
Contact: 021 689 9516

- 7) Addis in Cape**
Cuisine: Ethiopian
Location: City Bowl
Contact: 021 424 5722

Higher End
(Main meal - R100 and above)

1) *Die Wijnhuis

Cuisine: Mediterranean, Italian
Location: Newlands
Contact: 021 671 9705

2) *Barristers Grill

Cuisine: Grill and Seafood
Location: Newlands
Contact: 021 671 7907

3) Panama Jack's Taverna

Cuisine: Seafood
Location: Table Bay harbour
*Lunch rates are lower. For example
they offer a half-kilo of prawns for only
R60 during the week*
Contact: 021 448 1080

4) Olympia Cafe

Cuisine: Deli, Bakery, Coffee Shop
Location: Kalk Bay
Contact: 021 788 6396

5) *Bihari

Cuisine: Indian
Location: Newlands
Contact: 021 674 7186

6) Jonkershuis Constantia Eatery

Cuisine: Bistro
Location: Constantia
Contact: 021 794 4813

7) Moyo

Cuisine: African
Location: Kristenbosch National
Botanical Gardens
Contact: 021 762 9585

Internet Access

Most hotels will have access otherwise ask for directions to your nearest internet café.

Electricity

Voltage: 220/230 V

Adapter: You will need an adaptor for Plug M and sometimes plug C. Plug C is the two-pin plug commonly used in Europe.

Money

Withdrawals: We suggest that you use the campus ATM machines. They are situated on Middle Campus (next to the cafeteria), and Upper Campus (ground floor of the Leslie Social Science building and next to the library).

Credit Cards: When paying by credit card, we suggest that you ask vendors to swipe the card in your presence.

Exchange Rate: The current exchange rate is approximately 10.9 South Africa Rand to the US-Dollar.

Health and Emergencies

On campus:

- 1) Campus Protection Services: 021 650 2222/3
- 2) UCT Emergency Controller: 021 650 2175/6

Off Campus

- 1) Kingsbury Hospital (Wilderness Road, Claremont): 021 670 4000
- 2) Constantiaberg Medi-Clinic Hospital (Burnham Road, Plumstead): 021 799 2911 / 021 799 2196 (Emergency number)
- 3) Kenilworth Medicross (67 Rosmead Avenue, Kenilworth): 021 670 7640 – for doctor's visits

State Emergency Number (Police and Ambulance Services): **10111**

Private Ambulance Services: Netcare911: **082 911**
