J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

INTRODUCTION TO EVALUATIONS

A Randomized Evaluation is a type of Impact Evaluation that uses random assignment to allocate resources, run programs, or apply policies as part of the study design. Like all *impact evaluations*, the main purpose of randomized evaluations is to determine whether a program has an impact, and more specifically, to quantify *how large* that impact is. Impact evaluations measure program effectiveness typically by comparing outcomes of those (individuals, communities, schools, etc.) who received the program against those who did not. There are many methods of doing this. But randomized evaluations are generally considered the most rigorous and, all else equal, produce the most accurate (i.e. unbiased) results.

The methodology section covers the what, why, who, when, and how of randomized evaluations.

For more resources on randomized evaluations, see:

Running Randomized Evaluations
R. Glennerster and K. Takavarasha, November 2013

Field Experiments: Design, Analysis and Interpretation
A. Gerber and D. Green, May 2012

Evaluating Social Programs: Executive Education at J-PAL
Lecture videos and course content from a past **Executive Education Course**
A free **online version** of the course

Using Randomization in Development Economics Research: A Toolkit
E. Duflo, M. Kremer and R. Glennerster

Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons
M. Kremer

Field Experiments in Development Economics
E. Duflo, January 2006

Use of Randomization in the Evaluation of Development Effectiveness
E. Duflo and M. Kremer, July 2003

Scaling Up and Evaluation
E. Duflo, May 2003

Nonexperimental Versus Experimental Estimates of Earnings Impacts
S. Glazerman, D. Levy and D. Myers, May 2003

Impact Evaluation in Practice
P. Gertler, S. Martinez, P. Premand, L. Rawlings and C. Vermeersh

TABLE OF CONTENTS

1. WHAT IS EVALUATION AND WHY EVALUTE?

1.A. WHY EVALUATE?

The purpose of evaluation is not always clear, particularly for those who have watched surveys conducted, data entered, and then the ensuing reports filed away only to collect dust. This is most common when evaluations are imposed by others.

If, on the other hand, those responsible for the day-to-day operations of a program have critical questions, evaluations can help find answers. As an example, the NGO responsible for distributing chlorine pills may speak with their local field staff and hear stories of households diligently using the pills, and occasionally see improvements in their health. But each time it rains heavily, the clinics fill up with people suffering from diarrheal diseases. The NGO might wonder, "if people are using chlorine to treat their water, why are they getting sick when it rains? Even if the water is more contaminated, the chlorine should kill all the bacteria." The NGO may wonder whether the chlorine pills are indeed effective at killing bacteria. Are people using it in the right proportion? Maybe our field staff is not telling us the truth. Perhaps the intended beneficiaries are not using the pills. Perhaps they aren't even receiving them. And then when confronted with this fact, the field staff claims that during the rains it is difficult to reach households and distribute pills. Households, on the other hand, will reply that they most diligently use pills during the rains, and that the pills have helped them substantially.

Speaking to individuals at different levels of the organization as well as to stakeholders can uncover many stories of what is going on. These stories can be the basis for theories. But plausible explanations are not the same thing as answers. Evaluations involve developing hypotheses of what's going on, and then testing those hypotheses.

1.B. WHAT IS EVALUATION?

The word "evaluation" can be interpreted quite broadly. It means different things to different people and organizations. Engineers, for example, might evaluate or *test* the quality of a product design, the durability of a material, efficiency of a production process, or the safety of a bridge. Critics evaluate or *review* the quality of a restaurant, movie or book. A child psychologist may evaluate or *assess* the decision-making process of toddlers.

The researchers at J-PAL evaluate social programs and policies designed to improve the well-being of the world's poor. This is known as program evaluation.

Put simply, a program evaluation is meant to answer the question, "how is our program or policy doing?" This can have different implications depending on **who** is asking the question, and to whom they are talking. For example, if a donor asks the NGO director "how is our program doing?" she may imply, "have you been wasting our money?" This can feel interrogatory. Alternatively, if a politician asks her constituents, "how is our program doing?" she could imply, "is our program meeting your needs? How can we make it better for you?" Program evaluation, therefore, can be associated with positive or negative sentiments, depending on whether it is motivated by a *demand for accountability* versus a *desire to learn*.

J-PAL works with governments, NGOs, donors, and other partners who are more interested in learning the answer to the questions: How effective is our program? This question can be answered through an impact evaluation. There are many methods of doing impact evaluations. But the one used by J-PAL is the randomized evaluation.

At a very basic level, randomized evaluation can answer the question: *was the program effective? But if thoughtfully designed and implemented, it can also answer the questions, how effective was it? Were there unintended side-effects? Who benefited most? Who was harmed? Why did it work or not work? What lessons can be applied to other contexts, or if the program was scaled up? How cost-effective was the program?*

How does it compare to other programs designed to accomplish similar goals? To answer these (just as interesting, if not more interesting) questions, the impact evaluation should be part of a larger package of evaluations and exercises. Following the framework on *comprehensive evaluations* offered by Rossi, Freeman, and Lipsy, this package is covered in the subsequent sections:

1. Needs Assessment
2. Program Theory Assessment
3. Process Evaluation
4. Impact Evaluation
5. Cost-Benefit, Cost-Effectiveness, and Cost-Comparison Analysis
6. Goals, Outcomes, and Measurement

The first two assessments (Needs and Program Theory) examine what needs the program or policy is trying to fill and what are the steps by which it will achieve these objectives. Ideally, these steps should be formally set out by those implementing the program, before an impact evaluation is set up.

Process evaluations are useful for program managers and measure whether the milestones and deliverables are on schedule. Many organizations have established systems to track processes—often classified as Monitoring and Evaluation (M&E).

Impact evaluations are designed to measure whether programs or policies are succeeding in achieving their goals.

Lastly, Cost-benefit and Cost-effectiveness analyses are useful for the larger policy implications of a program. The first looks at whether the benefits achieved by the program are worth the costs. The second compares the benefits of this program to that of programs designed to achieve similar goals.

In conducting any assessment, evaluation, or analysis, it is imperative to think about how progress can be measured. Measuring indicators of progress – keeping the programs' goals and expected outcomes in mind—requires significant thought as well as a system of data collection. This is covered in Goals, Outcomes and Measurement.

1.B.1. NEEDS ASSESSMENT

Programs and policies are introduced to address a specific need. For example, we may observe that the incidence of diarrhea in a community is particularly high. This might be due to contaminated food or water, poor hygiene, or any number of plausible explanations. A needs assessment can help us identify the source of the problem and those most harmed.

For example, the problem may be due to the runoff of organic fertilizer which is polluting the drinking water used by certain communities.

Needs assessment is a systematic approach to identifying the nature and scope of a social problem, defining the target population to be served, and determining the service needed to meet the problem.

A needs assessment is essential because programs will be ineffective if the services are not properly designed to meet the need or if the need does not actually exist. So, for example, if the source of pollution contaminating drinking water is agricultural, investment in sanitation infrastructure such as toilets and sewage systems may not solve the problem. Needs assessments may be conducted using publicly available social indicators, surveys and censuses, interviews, etc.

1.B.2. PROGRAM THEORY ASSESSMENT

Social programs or policies are introduced to meet a social need. Meeting that need usually requires more thought than finding and pressing a single magic button, or taking a pill. For policymakers, it requires identifying the reasons that are causing undesirable outcomes (see Needs Assessment), and choosing a strategy from a large set of options to try to bring about different outcomes.

For example, if people are drinking unclean water, one program might be designed to prevent water from becoming contaminated—by improving sanitation infrastructure—while another may be designed to treat contaminated water using chlorine. One proposed intervention might target those responsible for the pollution. Another might target those who drink the water. One strategy may rest on the assumption that people don't know their water is dirty, another, that they are aware but have no access to chlorine, and even another, that despite awareness and access, people choose not to chlorinate their water for other reasons (e.g. misinformation, taste, cost, etc.). These programs must simultaneously navigate the capacity constraints (financial, human, and institutional) and political realities of their context. In conceiving an appropriate response policymakers implicitly make decisions about what is the best approach, and why. When this mental exercise is documented explicitly in a structured way, policymakers are conducting what can be called a *program theory assessment,* or *design assessment*.

A Program Theory Assessment models the theory behind the program, presenting a plausible and feasible plan for improving the target social condition. If the goals and assumptions are unreasonable, then there is little prospect that the program will be effective. Program theory assessment involves first articulating the program theory and then assessing how well the theory meets the targeted needs of the population. The methodologies used in program theory assessment include the *Logical Framework Approach* or Theory of Change.

The following table is a simple example of a log frame (logical framework):

Needs	Input	Output	Outcome	Impact	Long-Term Goal
People are frequently sick from drinking contaminated water and do not currently use methods to treat their water	NGO purchases chlorine tablets and develops infrastructure for distribution to households	Households receive chlorine tablets	Individuals stop drinking contaminated water and start drinking treated water	Incidence of diarrhea decreases	Decrease in mortality, particularly child mortality. Improved physical and cognitive development

1.B.3. PROCESS EVALUATION

Before it is ever launched, a program exists in concept—as a design, description or plan (see Program Theory Assessment). But once launched, the program meets on-the-ground realities: Is the organization adequately staffed and trained? Are responsibilities well assigned? Are the intermediate tasks being completed on schedule? If the program is designed to provide chlorine tablets to households to treat unclean water, for example, does the right number of chlorine tablets reach the appropriate distribution centers on time?

Process evaluation, also known as *implementation assessment* or *assessment of program process*, analyzes the effectiveness of program operations, implementation, and service delivery. When process evaluation is ongoing it is called *program monitoring* (as in Monitoring and Evaluation: M&E). Process evaluations help us determine, for example:

- Whether services and goals are properly aligned.
- Whether services are delivered as intended to the appropriate recipients.
- How well service delivery is organized.
- The effectiveness of program management.
- How efficiently program resources are used.[1]

Process evaluations are often used by managers as benchmarks to measure success, for example: the distribution of chlorine tablets is reaching 80% of the intended beneficiaries each week. These benchmarks may be set by program managers, and sometimes by donors. In many larger organizations, monitoring progress is the responsibility of an internal Monitoring and Evaluation (M&E) department. In order to determine whether benchmarks are being met, **data collection** mechanisms must be in place.

[1] Rossi, Peter, et al. Evaluation. A Systematic Approach. Thousand Oaks: Sage Publications, 1999.

1.B.4. IMPACT EVALUATION

Programs and policies are designed to achieve a certain goal (or set of goals). For example, a chlorine distribution program may be implemented specifically to combat high-incidence of waterborne illness in a region. We may want to know whether this program is succeeding in its goal. This isn't the same thing as asking, "Does chlorine kill bacteria?" or "Is the consumption of chlorine harmful?" Those questions can be answered in a *real* laboratory. For our program to achieve its goal of stopping illness, money must be allocated, tablets must be purchased, distribution mechanisms must be put in place, households must receive the tablets, households must use the tablets, and households must not consume untreated water. A program evaluation helps us determine whether all of these requirements are being met and if our goal is actually being achieved as intended.

As a normal part of operations, e.g. basic bookkeeping, certain information is produced, such as how many boxes of chlorine tablets have been shipped. This type of information can be used for process evaluation. But it cannot tell us whether we've successfully reduced the incidence of diarrhea. To measure impact, we must use more direct indicators such as the number of people who report suffering from diarrhea in the last two months.

Impact evaluations gauge the success of a program—where success can be broadly or narrowly defined. They help us weed out less effective interventions from successful ones and also help us improve existing programs.

Impact Evaluation

The primary purpose of impact evaluation is to determine whether a program has an impact (on a few key outcomes), and more specifically, to quantify *how large* that impact is. What is impact? In our chlorine example**, impact** is how much healthier people are because of the program than they would have been without the program. Or more specifically, how much lower the incidence of diarrhea is than it would have been otherwise.

Getting this number correct is more difficult than it sounds. It is possible to measure the incidence of diarrhea in a population that received the program. But "how they would have been otherwise" is impossible to measure directly—just as it is impossible to measure the United States economy today had the Nazis won World War II, or to determine today's most deadly disease if penicillin was not discovered in Alexander Fleming's dirty laboratory in 1928 in London. It is possible that Germany would have become the dominant economy in the world, or alternatively, the Nazis may have fallen just a few years later. It is possible that minor wounds would still be one of the largest killers, or alternatively, some close relative of penicillin could have been discovered in another laboratory in a different part of the world. In our chlorine example, it is possible that without chlorine, people would have remained just as sick as they were before. Or it is possible that they would have started boiling their water instead, and the only thing chlorine did was substitute one technology for another—suggesting that people are not really any healthier because of the chlorine.

Impact evaluations estimate program effectiveness usually by comparing outcomes of those (individuals, communities, schools, etc.) who participated in the program against those who did not participate. The key challenge in impact evaluation is finding a group of people *who did not participate*, but closely resemble the participants *had those participants not received the program.* Measuring outcomes in this comparison group is as close as we can get to measuring "how participants would have been otherwise." There are many methods of doing this and each method comes with its own assumptions.

A table comparing the different methodologies can be found in the Why Randomize section.

1.B.5. COST-BENEFIT/EFFECTIVENESS/COMPARISON ANALYSES

Two organizations may come up with very different strategies to tackle the same problem. If a community's water supply, for example, was contaminated leading to a large incidence of diarrhea, one NGO may advocate for investments in modern water and sanitation infrastructure, including a sewage system, piped water, etc. Another NGO may propose a distribution system where households are given free chlorine tablets to treat their own water at home. If these two methods were shown to be equally effective—each reducing diarrhea incidence by 80%, would local policymakers be just as happy implementing one versus the other? Probably not. They would also need to consider the cost of each strategy.

It is highly likely that modern infrastructure investments in an otherwise remote village would be prohibitively expensive. In this case, the choice may be clear. However, the options are not always so black and white. A more realistic (but still hypothetical) choice would be between an infrastructure investment that reduces diarrhea by 80% versus a chlorine distribution program that costs 1/100th the price, and reduces diarrhea by 50%.

A **cost-benefit analysis** quantifies the benefits and costs of an activity and puts them into the same metric (often by placing a monetary value on benefits). It attempts to answer the question: Is the program producing sufficient benefits to outweigh the costs? Or in other words, is society richer or poorer after making this investment? Trying to quantify the benefit of children's health in monetary terms, however, can be extremely difficult and subjective. Hence, when the exact value of the benefit lacks widespread consensus, this type of analysis may produce results that are more controversial than illuminating. This approach is most useful when there are multiple types of benefits and agreed ways of monetizing them.

A **cost-effectiveness analysis** takes the impact of a program (e.g. percent reduction in the incidence of diarrhea), and divides that by the cost of the program, generating a statistic such as: the number of cases of diarrhea prevented per dollar spent. This makes no judgment of the value of reducing diarrhea.

Lastly, a **cost comparison analysis** will take multiple programs and compare them using the same unit—allowing policy makers to ask: per dollar, how much does each of these strategies reduce diarrhea?

See the paper on "Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education" for more information.

1.B.6. GOALS, OUTCOMES AND MEASUREMENT

When conducting a program evaluation, governments and NGOs are often asked to distill a program's mission down to a handful of *outcomes* that, it is understood, will be used to define success. Adding to this difficulty, each outcome must be further simplified to an *indicator* such as the response to a survey question, or the score on a test.

More than daunting, this task can appear impossible and the request, absurd. In the process, evaluators can come across as caring only about data and statistics—not the lives of the people targeted by the program.

For certain goals, the corresponding indicators naturally follow. For example, if the goal of distributing chlorine tablets is to reduce waterborne illness, the related outcome may be a *reduction in diarrhea.* The corresponding indicator, *incidence of diarrhea*, could come from one question in a household survey where respondents are asked directly, "Has anyone in the household suffered from diarrhea in the past week?"

For other goals, such as "empowering women," or "improving civic mindedness" the outcomes may not fall as neatly into place. That doesn't mean that most goals are immeasurable. Rather, more thought and creativity must go into devising their corresponding indicators. For an example of difficult-to-measure outcomes, see **article**.

2. WHAT IS RANDOMIZATION AND WHY RANDOMIZE?

2.A. WHAT IS RANDOMIZATION?

In its most simple sense, randomization is what happens when a coin is flipped, a die is cast, or a name on a piece of paper is drawn blindly from a basket, and the outcome of that flip, cast, or draw determines what happens next. Perhaps, the outcome of the coin flip determines who has to do some chore; the role of the die determines who gets a pile of money; the draw of a name determines who gets to participate in some activity, or a survey. When these tools (the coin, the die, the lottery) are used to make decisions, the outcome can said to be left to chance, or, *randomized*.

Why do people let chance determine their fate? Sometimes, because they perceive it as fair. Other times, because uncertainty adds an element of excitement. Statisticians use randomization because, when enough people are *randomly chosen* to participate in a survey, conveniently, the attributes of those chosen individuals are *representative* of the entire group from which they were chosen. In other words, what is discovered about them is probably true about the larger group. Using a lottery to get a representative sample is known as *random sampling* or *random selection*.

When *two* groups are randomly selected from the same population, they *both* represent the larger group. They are not only *statistically equivalent* to the larger group; they are also statistically equivalent to each other. The same logic carries forward if more than two groups are randomly selected. When two or more groups are selected in this way, we can say that individuals have been *randomly assigned* to groups. This is called *random assignment*. (Random assignment is also the appropriate term when *all* individuals from the larger group divided randomly into different groups. As before, all groups *represent* the larger group and are statistically equivalent to each other.) *Random assignment* is the key element of randomized evaluation.

What happens next in a simple randomized evaluation (with two groups) is that one group receives the program that is being evaluated and the other does not. If we were to evaluate a water purification program using this method, we would randomly assign individuals to two groups. At the beginning, the two groups would be statistically equivalent (and are expected to have equivalent trajectories going forward). But then we introduce something that makes them different. One group would receive the water purification program and the other would not. Then, after some time, we could measure the relative health of individuals in the two groups. Because they were statistically equivalent at the beginning, any differences seen later on can be attributed to one having been given the water purification program, and the other not.

Why this method is used is covered in the **Why Randomize** section.

Randomized Evaluations go by many names:

- Randomized Controlled Trials
- Social Experiments
- Random Assignment Studies
- Randomized Field Trials
- Randomized Controlled Experiments

Randomized Evaluations are part of a larger set of evaluations called **Impact Evaluations**. Randomized evaluations are often deemed the gold standard of impact evaluation, because they consistently produce the most accurate results.

Like all *impact evaluations*, the primary purpose of randomized evaluations is to determine whether a program has an impact, and more specifically, to quantify *how large* that impact is. Impact evaluations measure program effectiveness typically by comparing

outcomes of those (individuals, communities, schools, etc.) who participated in the program against those who did not participate. There are many methods of doing this.

What distinguishes randomized evaluations from other non-randomized impact evaluations is that participation (and non-participation) is determined *randomly*—before the program begins. This *random assignment* is the method used in clinical trials to determine who gets a drug versus who gets a placebo when testing the effectiveness (and side-effects) of new drugs. As with clinical trials, those in the impact evaluation who were *randomly assigned* to the "treatment group" are eligible to receive the treatment (i.e. the program). And they are compared to those who were randomly assigned to the "control group" –those who do not receive the program. Because members of the groups (treatment and control) do not differ systematically at the outset of the experiment, any difference that subsequently arises between them can be attributed to the treatment rather than to other factors. Relative to results from non-randomized evaluations, results from randomized evaluations are:

- Less subject to methodological debates
- Easier to convey
- More likely to be convincing to program funders and/or policymakers

Beyond quantifying the intended outcomes caused by a program, randomized evaluations can also quantify the occurrence of unintended side-effects (good or bad). And like other methods of impact evaluation, randomized evaluations can also shed light on why the program has or fails to have the desired impact.

1. Randomization In the Context of "Evaluation"
Randomized evaluations are a type of impact evaluation that use a specific methodology for creating a comparison group—in particular, the methodology of random assignment. Impact evaluations are program evaluations that focus on measuring the final goals or outcomes of a program. There are many types of evaluations that can be relevant to programs—beyond simply measuring effectiveness. (See **What is Evaluation?**)

2. Methodology of Randomization
To better understand how the methodology works, see **How to conduct a randomized evaluation**.

2.B. WHY RANDOMIZE?

What is impact? In our chlorine example, impact is how much healthier people are because of the program than they would have been without the program. Or more specifically, it is how much lower the incidence of diarrhea is than it would have been otherwise.

Getting this number correct is more difficult than it sounds. It is possible to measure the incidence of diarrhea in a population that received the program. But "how they would have been otherwise" (termed, the *counterfactual*) is impossible to measure directly, it can only be inferred.

Constructing a Comparison Group

Impact evaluations estimate program effectiveness usually by comparing outcomes of those (individuals, communities, schools, etc.) who participated in the program against those who did not participate. The key challenge in impact evaluation is finding a group of people who did not participate, but closely resemble the participants, and in particular, the participants *if they had not received the program*. Measuring outcomes in this comparison group is as close as we can get to measuring "how participants would have been otherwise." Therefore, our estimate of impact is only as good as our comparison group is equivalent.

There are many methods of creating a comparison group. Some methods do a better job than others. All else equal, randomized evaluations do the best job. They generate a *statistically identical* comparison group, and therefore produce the most accurate

(unbiased) results. Or stated more strongly: other methods often produce misleading results—results that would lead policymakers to make exactly the opposite decision relative to where the truth would have directed them.

These other methods don't *always* give us the wrong answer, but they rely on more assumptions. When the assumptions hold, the answer is unbiased. But it is usually impossible, and always difficult, to ensure that the assumptions are true. In fact, it is likely that most debates about the validity of an evaluation are fueled by disagreements over whether these assumptions are reasonable.

Beyond escaping debates over assumptions, randomized evaluations produce results that are very easy to explain. A table comparing common methods of evaluation can be found here.

2.C. WHEN TO CONDUCT A RANDOMIZED EVALUATION?

The value added by rigorously evaluating a program or policy changes depending on when in the program or policy life cycle the evaluation is conducted. The evaluation should not come too soon: when the program is still taking shape, and kinks are being ironed out. And the evaluation should not come too late: after money has been allocated, and the program, rolled out, so that there is no longer space for a control group.

An ideal time is during the pilot phase of a program or before scaling up. During these phases there are often important questions that an evaluator would like to answer: How effective is the program? Is it effective among different populations? Are certain aspects are working better than others, and can "the others" be improved? Is it effective when it reaches a larger population?

During the pilot phase, the effects of a program on a particular population are unknown. The program itself may be new or it may be an established program that is targeting a new population. In both cases program heads and policymakers may wish to better understand the effectiveness of a program and how it might be improved. Almost by definition, the pilot program will reach only a portion of the target population, making it possible to conduct a randomized evaluation. After the pilot phase, if the program is shown to be effective, leading to increased support, and in turn, more resources allocated, it can be replicated or scaled up to reach the remaining target population.

One example of a well-timed evaluation is that of PROGRESA, a conditional cash transfer program in Mexico launched in 1997. The policy gave mothers cash grants for their family as long as they ensured their children attended school regularly and received scheduled vaccinations. The political party, which had been in power for the prior 68 years, the Institutional Revolutionary Party (PRI), was facing inevitable defeat in the up-coming elections. A probable outcome of electoral defeat was the dismantling of incumbent programs such as PROGRESA. To build support for the program's survival, PRI planned to clearly demonstrate the policy's effectiveness in improving child health and education outcomes.

PROGRESA was first introduced as a pilot program in rural areas of seven states. Out of 506 communities sampled by the Mexican government for the pilot, 320 were randomly assigned to treatment and 186 to the comparison. Comparing treatment and control groups after one year, it was found to successfully improve these child-level outcomes. As hoped, the program's popularity expanded from its initial supporters and direct beneficiaries to the entire nation.

Following the widely-predicted defeat of PRI in the 2000 elections, the new political party, PAN took power and inherited an immensely popular program. Instead of dismantling PROGRESA, PAN changed the program's name to OPORTUNIDADES, and expanded it nation-wide.

The program was soon replicated in other countries, such Nicaragua, Ecuador, and Honduras. And following Mexico's lead, these new countries conducted pilot studies to test the impact of PROGRESA-like programs on their populations before scaling up.

2.D. WHEN IS A RANDOMIZED EVALUATION NOT APPROPRIATE?

Randomized evaluations may not be appropriate:

1. **When evaluating macro policies.**

No evaluator has the political power to conduct a randomized evaluation of different monetary policies. One could not randomly assign a floating exchange rate to Japan and other nations and a fixed exchange rate to the United States and a different group of nations.

2. **When it is unethical or politically unfeasible to deny a program to a control group.**

It would be unethical to deny a drug whose benefits have already been documented to some patients for the sake of an experiment if there are no resource constraints.

3. If the program is changing during the course of the experiment.

If midway through an experiment, a program changes from providing a water treatment solution to providing a water treatment solution and a latrine, it will be difficult to interpret what part of the program produced the observed results.

4. If the program under experimental conditions differs significantly from how it will be under normal conditions.

During an experiment participants may be more likely to use a water treatment solution if they are encouraged or given incentives. In normal conditions, without encouragement or incentives, fewer people may actually use the water treatment solution even if they own it and know how to use it.

As a caveat, this type of evaluation may be valuable in testing a proof of concept. It would simply be asking the question, "can this program or policy be effective?" It would not be expected to produce generalizable results.

5. If a RCT is too time-consuming or costly and therefore not cost-effective.

For example, due to a government policy, an organization may not have sufficient time to pilot a program and evaluate it before rolling it out.

6. If threats such as attrition and spill-over are too difficult to control for and hurt the integrity of the experiment.

An organization may decide to test the impact of a deworming drug on school attendance at a particular school. Because deworming drugs have a spill-over effect (the health of one student impacts the health of another), it will be difficult to accurately measure the impact of the drug. In this case, a solution could be to randomize at a school level rather than at a student level.

7. If sample size is too small.

If there are too few subjects participating in the pilot, even if the program were successful, there may not be enough observations to statistically detect an impact.

3. HOW TO CONDUCT A RANDOMIZED EVALUATION?

Some refer to randomized evaluations as the *gold standard* of impact evaluation because they are inarguably the most rigorous—meaning they require the fewest assumptions, or leaps of faith, when drawing conclusions from the results. Being the most rigorous, however, does not by itself mean they require significantly more work or cost. In fact, frontloading the work by randomizing to ensure equivalent groups at the beginning (see What is Randomization and Why Randomize) can reduce the amount of statistical work synthesizing an equivalent comparison group later on in the analysis phase.

There are certainly challenges with conducting a randomized evaluation: convincing program implementers to randomize, thinking about the appropriate evaluation design, ensuring that the integrity of the evaluation design (random assignment) is maintained. But the bulk of the work and cost come from ensuring a sufficient sample to detect impact (a prerequisite of even non-randomized evaluations) and figuring out why the program works or does not work.

3.1 PLANNING AN EVALUATION

In planning an evaluation it is important to identify key questions the organization may have. From these, we can determine how many of those questions can be answered from prior impact evaluations or from an improved systems of **process evaluation**. Assuming we haven't found all our answers, we must then pick a few top priority questions that will be the primary focus of our impact evaluation. Finally we should draw up plans to answer as many questions as we can, keeping in mind that fewer high quality impact studies are more valuable than many poor quality ones.

The first step in an evaluation is to revisit the program's goals and how we expect those goals to be achieved. A logical framework or theory of change model can help in this process. (See Program Theory Assessment) As part of assessing the purpose and strategy of a program, we must think about key outcomes, the expected pathways to achieve those outcomes, and reasonable milestones that indicate we're traveling down the right path. As expected in an evaluation, these outcomes and milestones will need to be measured, and therefore transformed into "indicators" and ultimately "data". (See Goals, Outcomes, and Measurement.)

Only after we have a good sense of the pathways, the scope of influence, and a plan for how we will measure progress, can we think about the actual design of the evaluation.

3.2 HOW TO DESIGN AN EVALUATION?

An evaluation design requires a considerable amount of thought. First comes the conceptual pieces: what do we plan to learn from this evaluation? What are the relevant questions? What outcomes are expected? How can they be measured?

Next, come the design questions:

- What is the appropriate level or unit of randomization?
- What is the appropriate method of randomization?
- Beyond the political, administrative and ethical constrains, what technical issues could compromise the integrity of our study, and how can we mitigate these threats in the design?
- How would we implement the randomization?
- What is the necessary sample size to answer our questions? (How many people do we need to include in the study, both as participants, but also as survey respondents?)

1. Unit of Randomization

In designing our evaluation we must decide at what level we will randomize: what unit will be subject to random assignment? Will it be individuals or groupings of individuals, such as households, villages, districts, schools, clinics, church groups, firms, and credit associations? (When we randomize groups of individuals—even though we care about and measure individual outcomes—this is referred to as a *cluster randomized trial*.) For example, if we managed to secure enough chlorine pills for one thousand households to treat contaminated water (out of, say, ten thousand households who use the same contaminated source of drinking water), do we expect to randomly assign *households* to the treatment and control groups? This means that some households will be given chlorine pills, but some of their immediate neighbors will be denied chlorine pills. Is that feasible? Ethical?

For this type of program, it probably wouldn't be feasible to randomize at an even smaller unit than the household, for example the individual level. It would imply that some children within a household are given chlorine pills and some of their siblings are not. If all household members drink from the same treated tank of water, individual randomization would be physically impossible, regardless of the ethical considerations.

Perhaps the appropriate unit of randomization is the community, where some communities will receive chlorine, other communities will not, but within a "treatment" community all households (implying all neighbors) are eligible to receive the pills. There are many things to consider when determining the appropriate level of randomization, of which ethics and feasibility are only two. Seven considerations are listed below.

1. What unit does the program target for treatment?
2. What is the unit of analysis?
3. Is the evaluation design fair?
4. Is a randomized evaluation politically feasible?
5. Is a randomized evaluation logistically feasible?
6. What spillovers and other effects will need to be taken into account?
7. What sample size and power do we require to detect effects of the program?

1. *What unit does the program target for treatment:* If chlorine tablets are meant to be dissolved in water storage tanks that in our region all households typically already own, then some households could be selected to receive chlorine, and others not. In this case, the unit of randomization would be at the household level. However, if the storage tank is typically located outside and used by a cluster of households, then it would be impossible to randomly assign some households in that cluster to the control group—they all drink the same (treated) water as the treatment households. Then, the most natural unit of randomization may be the "clusters of households" that use a common water tank.

2. *What is the unit of analysis:* If the evaluation is concerned with community level effects then the most natural level of randomization is probably the community. For example, imagine our outcome measure is incidence of "hospitalization" due to diarrhea, and it is most economical to measure this using administrative records at community clinics, and furthermore, those records remain anonymous. We would not be able to distinguish whether people who were hospitalized were from treatment households or control households. However, if the entire community is in the treatment group, we could compare the records from clinics in treatment communities against those of control communities.

3. *Fairness:* The program should be perceived as fair. If I've been denied chlorine pills, but my immediate neighbors receive them, I might be angry with my neighbors, angry with the NGO, and I might be less willing to fill out some questionnaire on chlorine usage when surveyors knock at my door. And the NGO might not be enthusiastic about upsetting its community members. On the other hand, if my entire community didn't get it, but a neighboring community did, I might never hear of their program, so have nothing to complain about; or I could think that this was just a village-level choice and my village chose not to invest. Of course, people may be equally upset about a community-level design. We could try to expand the unit of

randomization, or think of other strategies to mitigate people's dissatisfaction. The fact that everyone is not helped may be unfair. (See ethical issues.) But given that we cannot help everyone (usually due to capacity constraints), and our desire to improve and evaluate, how can we allocate in a way that simultaneously allows us to create an equivalent control group, and is seen as fair by the people we're trying to help.

4. *Political Feasibility:* It may not be feasible politically to randomize at the household level. For example, a community may demand that all needy people receive assistance, making it impossible to randomize at the individual or household level. In some cases, a leader may require that all members of her community receive assistance. Or she may be more comfortable having a randomly selected half be treated (with certainty) than risk having no one treated (were her village assigned to the control group). In one case she may comply with the study and in another, she may not.

5. *Logistical Feasibility:* Sometimes it is logistically impossible to ensure that some households remain "control households". For example, if chlorine distribution requires hiring a merchant within each village, setting up a stall where village members pick up their pills, it may be inefficient to ask the distribution agent to screen out control households. It could add bureaucracy, waste time, and distort what a real program should actually look like. Or even if the merchant could easily screen, households may simply share the pills with their "control group neighbors". Then the control group would be impacted by the program, and no longer serve as a good comparison group. (Remember, the comparison group is meant to represent life without the program. In this case, it would make sense to randomize at the village level, and then simply hire merchants in treatment villages and not in control villages.

6. *Containing spillovers and other effects:* Even if it is feasible to randomize at the household level—to give some households chlorine tablets and not others—it may not be feasible to contain the impact within just the treatment households. If control group individuals are affected by the presence of the program—they benefit from fewer sick neighbors (spillover effects), or drink the water from treatment neighbors (don't comply with the random assignment, and cross over to the treatment group), they no longer represent a good comparison group.

7. *Sample size and power:* The ability to detect real effects depends on the sample size. When more people are sampled from a larger population, statistically, they better represent the population. For example, if we survey two thousand households, and randomize at the household level (one thousand treatment, one thousand control), we effectively have a sample size of two thousand households. But if we randomized at the village level, and each village has one hundred households, then we would have only ten treatment villages and ten control. In this case, we may be measuring diarrhea at the household level, but because we randomized at the village level, it is possible we have an effective sample size closer to ten (even though we are surveying two thousand households)! In truth, the effective sample size, could be anywhere from ten to two thousand, depending on how similar households within villages are to their fellow villagers. (See: **sample size**.) With an effective sample size of ten, we may not be able to detect real effects. This may influence our choice as to the appropriate level of randomization.

There are many considerations when determining the appropriate level of randomization. Evaluators cannot simply sit at a computer, press a button, produce a list, and impose an evaluation design on an organization from thousands of miles away. Evaluators must have a deep and broad understanding of the implementing organization, their program, and the context and work in partnership to determine the appropriate level of randomization given the particular circumstances.

2. Different Methods of Randomization

If my organization can secure one thousand chlorine pills per day, so I can treat one thousand out of an eligible two thousand households per day, I could choose to treat the same one thousand households in perpetuity. Alternatively I could rotate so that every other day, each household gets to drink clean water one of those days. I may feel that the latter option makes no sense. If everyone is drinking dirty water half the days, I may expect zero impact on anyone. So I may choose one thousand households that will receive the pills daily. If randomizing, I may perform a simple "lottery" to determine which thousand households get the pill: I write all two thousand names onto a small piece of paper, put those pieces of paper into a basket, shake the basket up, close my eyes and pull one thousand pieces of paper out. Intuitively, this would be called, a *lottery* design.

Alternatively, If I were to rotate households instead of every day, every year, and randomly assign the order in which they get treated, and then in one out of those two years households would be considered the treatment group, and in the other year, they would be part of the control group. If I were to measure outcomes at the end of each year, this would be called a *rotation* design.

Say I can secure five hundred pills per day this year, but next year I expect to one thousand per day, and the following year, two thousand per day. I could randomly choose five hundred households to get the pill in the first year, another five hundred to be added in the second year, and the remaining thousand get it in the third year. This would be called, a *phase-in* design.

There are seven possible randomization designs—the lottery design, phase-in design, rotation design, encouragement design, the varying levels of treatment design, and two-stage randomization. These designs are not necessarily mutually exclusive. Their advantages and disadvantages are summarized in the table below:

A table comparing strategies used to create randomized comparison groups can be found here.

3. Threats to the Design

a) Spillovers

A spillover effect occurs when a program intended to help targeted participants unintentionally impacts the comparison group as well (either positively or negatively). The comparison group is supposed to represent outcomes had the program not been implemented. If this comparison group has been touched by the program, its role mimicking the counterfactual is now compromised, and the ensuing impact measure may be biased. There are ways of mitigating spillover effects, for example, changing the level of randomization.

For example, one source of sickness may be drinking contaminated water. But another source is playing with neighboring children who are themselves sick. If I am in the control group, and the program treats my neighbors, and those neighbors are no longer sick, that reduces my chance of getting sick. And even though I may be in the control group, I have now been affected by the treatment of my neighbors. I would no longer represent a good comparison group. This is known as a spillover effect, in particular a positive spillover. To mitigate this, we could randomize at the community level. Doing so would mean that if our community were assigned to the control group, I and all of my neighbors would share the same status. I would be less likely to play with children from a different community and therefore less likely be impacted by the intervention. Or if assigned to the treatment group, we would not positively impact others.

(Of course, we may actually want to know how these spillovers occur, and design accordingly.)

b) Crossovers

Another possibility is that my household has been assigned to the control group, but my neighbor is in the treatment group, and so my mother knows their water is clean, and she sends me to their house to drink. In a sense, I am finding my way into the treatment group, even though I was assigned to the control group. When people deliberately defy their treatment designation (knowingly or unknowingly), and as a result, outcomes are altered, this would be considered a crossover effect. As with spillovers, by crossing over I no longer represent a good comparison group—since I have clearly been affected by the existence of the program. As before, changing the level of randomization could mitigate crossover effects.

4. Mechanics of Randomization

Once the unit and method of randomization have been determined, it is time to randomly assign individuals, households, communities, or any unit to either the treatment or control group.

a) Simple Lottery

Generally to start with, we need a list of (individual, household head, or village) names. Next, there are several ways to proceed. We could write all names onto a small piece of paper, put those pieces of paper into a basket, shake the basket up, close our eyes and pull one thousand pieces of paper out. Those will make up the treatment group and the remainder could be the control group. (or vice versa) We may do this as part of a public lottery. Similarly, we could go down the list, one-by-one and flip a coin to determine treatment status. However, we don't always divide the study population exactly in half. We may wish to include 30

percent in the treatment group and 70 in the control. Or if we had a phase-in method with three periods, we may want to divide the population into three groups. Also very common, we will test multiple treatments at the same time—also requiring several groups. In these more sophisticated evaluation designs, a coin flip will not suffice.

Typically, we will write a computer program to randomly assign names to groups.

b) Spot-Randomization

Sometimes we do not have a list beforehand. For example, if individuals enter a clinic with symptoms of malaria, the decision of whether to administer the World Health Organization's standard "DOTS" treatment or an enhanced alternative must be made on-the-spot. The treatment could be determined by the nurse at the clinic using the flip of a coin. But we may be concerned that the nurse would ignore the random assignment if she has an opinion of which treatment is better and which patients are more "deserving" than others. Alternatives could include computerized or cell-phone based randomization.

c) Stratified Randomization

Frequently, the target population is divided into subgroups before randomizing. For example, a group of individuals can be divided into smaller groups based on gender, ethnicity, or age. Or villages could be divided into geographic regions. This division into subgroups before randomization is called stratification. Then the randomization exercise takes place within each of the strata (subgroups). This is done to ensure that the treatment and control groups have balanced proportions of treatment and control within each group. It is conceivable that with a small sample, we find that without stratifying, we end up with more females in our treatment group than males. The primary purpose of stratification is statistical and relates to sample size. The decision to stratify has no bearing on whether the results are biased.

5. Sample Selection and Sample Size

An experiment must be sensitive enough to detect outcome differences between the treatment and the comparison groups. The sensitivity of a design is measured by statistical power, which, among other factors, depends on the sample size – that is, the number of units randomly assigned and the number of units surveyed.

Once again, let's take our example of waterborne illness in a community. And let us assume that we have chosen to distribute chlorine tablets to households to test their impact on the incidence of diarrhea. But let us also assume that we only have a very limited budget for our test phase, and so we would like to minimize the number of households that are included in the survey while still ensuring that we can know whether any change in incidence is due to the chlorine tablets and not to random chance. How many households should receive the tablets, and how many should be surveyed? Is five households enough? 100? 200? How many households should be in the control group? Tests for statistical power help us answer these questions.

For more information on how to estimate the required sample size, see:

Duflo, Esther, Glennerster, Rachel, and Kremer, Michael, "Using Randomization in Development Economics Research: A Toolkit" (2006). MIT Department of Economics Working Paper No. 06-36.

Bloom, H.S. (1995): "Minimum Detectable Effects: A simple way to report the statistical power of experimental designs," Evaluation Review 19, 547-56.

3.3 WHO PARTICIPATES IN RANDOMIZED EVALUATIONS?

Each randomized evaluation (RE) is made possible through a partnership between researchers, organizations that run the programs to be evaluated (such as governments or NGOs), donors who fund the programs and evaluation, research centers who employ the staff associated with each evaluation, and research subjects who agree to participate. The social programs that the REs evaluate are often designed to target a certain population, for example, the poor or otherwise disadvantaged. The targeted populations of these programs are also the research subjects who participate in REs.

For an overview of major players conducting REs, click here.

The question of who participates in a randomized evaluations touches on some of the most sensitive issues faced by an evaluator. In answering this question an evaluator must consider what is ethical and fair. It would be unethical, for example, to deprive a household of a water treatment solution for the sake of an experiment if the household would have otherwise had access to the solution.

1. Ethical Issues

So how can an evaluator conduct an experiment while still meeting fair and ethical standards?

Randomized evaluations can be appropriate in situations when there are resource constraints. Typically an organization does not have a large enough budget to provide everyone in a community or district or country with a program. Because of budget constraints an organization must decide who receives the program and who does not. Even if they target the subgroup of people who particularly need the program, or those who would benefit most, they are unlikely to be able to cover everyone even in the target subgroups. This provides an evaluator with an opportunity to conduct a randomized evaluation. An evaluator can introduce an element of randomization into the decision of how to allocate scarce resources within the target subgroup.

An evaluator must not only ensure that an experiment is ethical, but that it is also fair. In randomly assigning participants to the control or experimental group, an evaluator should ensure that everyone has an equal chance of being in the experimental group and receiving the program. Methods of fairly selecting participants include using a lottery, phasing in a program, and rotating participants through the program to ensure that everyone benefits. The selection process should also be transparent and appear fair to the community.

Typically evaluators are faced with the problem of allocating a program that is clearly beneficial, such as deworming drugs, or a water treatment solution. In other words, the ethical dilemma surfaces when creating a group of individuals who will be denied the program. Sometimes, however, the benefits have not been proven, meaning it is possible the program could potentially make participants worse off. For example, drug companies often face this problem when testing new treatments on patients. In this case, an evaluator must put as much energy into ensuring that participants in the treatment group are not harmed. If there is any potential risk in participating, then everyone involved must be informed of the risks and give their consent to participate. Even if there do not appear to be risks, any experiment should get the informed consent of all participants (in both treatment and comparison groups). Human subject protocols have been developed by different nations and organizations and should be followed carefully. (See below)

2. Research Subjects and the Institutional Review Board

An Institutional Review Board (IRB), also known as an independent ethics committee or a human subjects review board, is a group that has been formally designated by an institution (such as a university or non-profit) to approve, monitor, and review research involving humans as participants. An IRB's objective is to assure, both in advance of and by periodic review, which appropriate steps are taken to protect the rights and welfare of humans participating as subjects in a research study.

Because J-PAL studies involve human participants, J-PAL's affiliates and their staff ensure that studies meet the guidelines of ethical research methods. This includes:

- Receiving institutional review board (IRB) approvals for each study before it begins,
- All study personnel completing an IRB training course,
- Adhering to the IRB approved research protocol and guidelines throughout the course of the study.

3.4 HOW TO IMPLEMENT?

Once an evaluation design has been finalized, the evaluator must remain involved to monitor data collection as well as the implementation of the intervention being evaluated. If respondents drop out during the data collection phase the results are susceptible to **attrition bias**, compromising their validity. Attrition is covered in this section. Other threats in the data collection phase such as poor measurement instruments, reporting bias, etc., are equally important, but are not covered here. For best practices on data collection see:

Deaton, A. (1997): The Analysis of Household Surveys. World Bank, International Bank for Reconstruction and Development

In the implementation of the intervention, the integrity of the randomization should remain intact. Unless intentionally incorporated into the study's design, *spillovers and crossovers* should be minimized, or at the very least, thoroughly documented.

1. Threats to Data Collection

a) Attrition

Attrition occurs when evaluators fail to collect data on individuals who were selected as part of the original sample. Note that the treatment and control groups, through random assignment, are constructed to be statistically identical at the beginning. The control group is meant to resemble the counterfactual—what would have happened to the treatment group had the treatment not been offered. (See: Why Randomize?). If individuals who drop out of the study are "identical" in both the treatment and control groups, meaning the depleted control group still represents a valid counterfactual to the depleted treatment group.This will reduce our sample size, and could truncate the target population to which our results can be generalized, but it will not compromise the "truth" of the results (at least as applied to the restricted population).

For example, suppose our study area is rural, and that many household members spend significant portions of the year working in urban areas. Further suppose we created our sample and collected baseline data when migrant household members were home during the harvests and incidentally available for our study. If we collect our endline data during off-peak season, the migrant family members will have returned to their city jobs and will be unavailable for our survey. Assuming these are the same people in both the treatment and control groups, our study will now be restricted to only non-migrants. If the non-migrant population in the control group represents a good counterfactual to the non-migrant population in the treatment group, our impact estimates will be perfectly valid—but only applicable to the non-migrant population.

However, if attrition takes a different shape in the two groups, and the remaining control group no longer serves as a good counterfactual, this could bias our results. Using our example of waterborne illness, suppose that in the control group more children and mothers are ill. As a result, the young men who typically migrate to the cities during off-peak seasons stay back to help the family. Households that were assigned to the control group contain more migrants during our endline. The baseline demographics of the treatment and control groups are now different (whereas originally, they were balanced). It is entirely feasible that these migrants, of peak working age, are typically healthier. Now, even though our treatment succeeded in producing healthier children and mothers on average, our control group contains more healthy migrant workers, on average. When measuring the incidence of diarrhea, outcomes of the healthy migrants in the control group could offset those of their sicker family members. Then, when comparing the treatment and control groups, we could see no impact at all and may conclude the treatment was ineffective. This result would be false and misleading.

In this simplified example, we could forcibly reintroduce balance between the comparison and experimental groups by removing all migrants from our sample. Frequently, however, characteristics that could dependably identify both real and would-be attritions (those who disappear) have not been measured, or are impossible to observe. Predicting attrition can be as difficult as

predicting participation in non-randomized trials. Similarly, attrition bias can be as damaging as selection bias when making causal inference.

2. Spillovers and Crossovers

Spillovers occur when individuals in the control group are somehow affected by the treatment. For example, if certain children are in the control group of a chlorine dispensing study, but play with children who are in the treatment group, they now have friends who are less likely to be sick, and are therefore less likely to become sick themselves. In this case, they are indirectly impacted by the program, even though they have been assigned to the control group. Individuals who "crossover" are control those who find a way to be directly treated. For example, if the mother of a control group child sends her child to drink from the water supply of a treatment group household, she is finding her way into the treatment group. Impartial compliance is a broader term that encapsulates crossovers, and also treatment individuals who deliberately choose not to participate (or chlorinate their water, in this example).

When a study suffers from spillovers and crossovers, in many circumstances it is still possible to produce valid results, using statistical techniques. But these come with certain assumptions—many of which we were trying to avoid when turning to randomization in the first place. For example, if spillovers can be predicted using observed variables, they can be controlled for. With impartial compliance, if we assume that those who did not comply were unaffected by the intervention, and by the same token, the individuals who crossed over were affected in the same way as those from the treatment group who were treated, we can infer the impact of our program. But as discussed in the Why Randomize section, the more assumptions we make, the less firm ground we stand on when claiming the intervention caused any measured outcomes.

3.5 HOW TO OBTAIN RESULTS?

At the end of an intervention (or at least the evaluation period for the intervention), endline data must be collected to measure final outcomes. Assuming the integrity of the random assignment was maintained, and data collection was well-administered, it is time to analyze the data. The simplest method is to measure the average outcome of the treatment group and compare it to the average outcome of the control group. The difference represents the program's impact. To determine whether this impact is statistically significant, one can test the equality of means, using a simple t-test. One of the many benefits of randomized evaluations is that the impact can be measured without advanced statistical techniques. More complicated analyses can be performed. For example, regressions controlling for other characteristics can be run to add precision. However, as the complexity of the analysis mounts, the number of potential missteps increases. Therefore, the evaluator must be knowledgeable and careful when performing such analyses.

It is worth noting that when a result is obtained, we have not uncovered the truth with 100 percent certainty. We have produced an estimate that is close to the truth with a certain degree of probability. The larger our sample size, (the smaller our standard errors will be and) the more certain we are. But we can never be 100 percent certain.

This fact leads to two very common pitfalls in analysis:

1) **Multiple Outcomes:** Randomization does not ensure the estimated impact is perfectly accurate. The measured impact is unbiased, but it is still an estimate. Random chance allows for some margin for error around the truth. Quite frequently the estimate will be extremely close to the truth. Occasionally, the estimate will deviate slightly more. Rarely, it will deviate significantly. If we look at one outcome measure, there is some chance that it has deviated significantly from the truth. But this is highly unlikely. If we look at many outcome measures, most will be close, but some will deviate. The more indicators we look at, the more likely one or more will deviate significantly. For example, assume the chlorine pills we distributed to fight waterborne illness in our water purification program were faulty or never used—if twenty outcome measures are compared, it is in fact very likely that one comparison will suggest significant improvement in health and one will indicate significant decline due to our program. So if we look at enough outcome measures eventually we will stumble upon one that is significantly different between the treatment and control groups. This is not a problem, per se. The problem arises when the evaluator "data mines,"

looking at outcomes until she finds a significant impact, reports this one result, and fails to report the other insignificant results that were discovered in the search.

2) **Sub-group analysis**: Similarly, just as an evaluator can data mine by looking at many different outcome measures, the evaluator can also dig out a significant result by looking at different sub-groups in isolation. For example, it might be that the chlorine has no apparent impact on household health as a whole. It may be reasonable to look at whether it has an impact on children within the household, or girls in particular. But we may be tempted to compare boys and girls of different age groups, of different compositions of households, in different combinations. We may discover that there is a significantly better health in the treatment group for the subgroup of boys between the ages of 6 and 8, who happen to have one sister, one grandparent living in the household, and where there the household owns a TV and livestock. We could even concoct a plausible story for why this subgroup would be affected and other subgroups not. But if we stumbled upon this one positive impact after finding a series of insignificant impacts for other subgroups, it is likely that the difference is due simply to random chance—not our program.

3.6 HOW TO DRAW POLICY IMPLICATIONS?

Having performed a perfect randomized evaluation, and an honest analysis of the results, with a certain level of confidence we can draw conclusions about how our program impacted this specific target population. For example: "Our chlorine distribution program caused a reduction in the incidence of diarrhea in children of our target population by 20 percentage points." This statement is scientifically legitimate, or *internally valid*. The rigor of our design cannot tell us, however, whether this same program would have the same or any impact if replicated in a different target population, or if scaled up. Unlike internal validity, which a well-conducted randomized evaluation can provide, *external validity*, or *generalizability*, is more difficult to obtain. To extrapolate how these results would apply in a different context, we need to depart from our scientific rigor, and begin to rely on assumptions. Depending on our knowledge of the context of our evaluation, and other contexts upon which we would like to generalize the results, our assumptions may be more or less reasonable.

However, the methodology we chose—a randomized evaluation—does not provide internal validity at the cost of external validity. External validity is a function of the program design, the service providers, the beneficiaries, and the environment in which the program evaluation was conducted. The results from any program evaluation are subject to these same contextual realities when used to draw inferences for similar programs or policies implemented elsewhere. What the randomized evaluation buys us is more certainty that our results are at least internally valid.

4. HISTORY OF RANDOMIZED EVALUATIONS?

To read about when randomized evaluations are appropriate, see: "When to conduct a randomized evaluation?" or "When is randomization (not) appropriate?"

4.A. HISTORY OF RANDOMIZED EVALUATIONS?

1. Clinical Trials

The concept of a control and experimental group was introduced in 1747 by James Lind when he demonstrated the benefits of citrus fruits in preventing scurvy using a scientific experiment.[1] As a result of his work, Lind is considered to be the father of clinical trials. The method of randomly assigning subjects to control and treatment groups, however, was not developed until the 1920s.

2. Agricultural Experiments

Randomization was introduced to scientific experimentation in the 1920s when Neyman and Fisher conducted the first randomized trials in separate agricultural experiments. Fisher's field experimental work culminated with his landmark book, The Design of Experiments, which was a main catalyst for the much of the growth of randomized evaluations.[2]

3. Social Programs

Randomized trials were introduced to government sponsored social experiments between 1960 and 1990. Rather than small-scale experiments conducted on plants and animals, these new social experiments were significantly larger in scale and focused on people as the subjects of interest. The idea of conducting social policy experiments grew out of a 1960s debate over the merits of the welfare system. The model of social experimentation was later applied both in Europe and the United States to evaluate other programs such as electricity pricing schemes, employment programs, and housing allowances. Since then social experiments have been used across disciplines and in a variety of settings around the world to guide policy decisions.[3]

The Abdul Latif Jameel Poverty Action Lab (J-PAL) was founded in June 2003 as a network of affiliated professors around the world who are united by their use of randomized evaluations to answer questions critical to poverty alleviation.

[1] Thomas, Duncan P. Sailors, Scurvy and Science. Journal of the Royal Society of Medicine. 90 (1997).
[2] Levitt, Steven D. and John A. List. 2009. "Field Experiments in Economics: The Past, The Present, and The Future." European Economic Review 53(1): 1-18.
[3] ibid

4.B. WHO CONDUCTS RANDOMIZED EVALUATIONS?

J-PAL was founded in 2003 as a network of affiliated professors who conduct impact evaluations using the randomized evaluation (RE) methodology to answer questions critical to poverty alleviation. J-PAL affiliates also conduct non-randomized research, and many other people and organizations conduct REs. For a brief history of RE's journey from clinical trials to agricultural experiments to social programs and poverty alleviation, click here.

Since J-PAL's founding, more than 200 organizations have partnered with a J-PAL affiliate on a RE. Amongst key players in poverty alleviation and development, the idea of REs is now fairly well-known.

Of the top ten *U.S. foundations,*[1] four of the six that work on international development have partnered with a J-PAL affiliate on a RE. These include the Bill & Melinda Gates Foundation, the Ford Foundation, the William and Flora Hewlett Foundation, and the John D. and Catherine T. MacArthur Foundation.[2]

Of the top ten *multilateral organizations*,[3] four have partnered with a J-PAL affiliate on a RE (the World Bank, the Asian Development Bank, Unicef, and the Inter-American Development Bank), and six of the ten have sent staff to J-PAL's training courses.

Of "The Big Eight" relief organizations,[4] Save the Children, Catholic Relief Services, CARE, and Oxfam have partnered with a J-PAL affiliate on a RE. The International Rescue Committee is doing REs on its own. And six of the eight have sent staff to J-PAL's training courses.

Governments also partner with J-PAL affiliates. Major donor country partners include the United States (USAID, MCC), France (Le Ministre de la Jeunesse et des Solidarités Actives), Sweden, and the United Kingdom (DFID). Developing country government partners have been both at the national level (e.g. Kenya's Ministry of Education and the Government of Sierra Leone's Decentralization Secretariat) and the sub-national level (e.g. the Government of Andhra Pradesh, the Gujarat Pollution Control Board, and the Rajasthan police).

A number of research centers have been established with the support or under the direction of J-PAL affiliates. These research centers often run affiliates' REs and employ the staff associated with each RE. These research centers include: Innovations for Poverty Action (IPA), Centre for Microfinance, Center for International Development's Micro-Development Initiative, Center for Effective Global Action, Ideas42, and the Small Enterprise Finance Center.

Private companies also conduct randomized evaluations of social programs. Mathematica Policy Research and Abt Associates are two examples.

[1] When measured by endowment.
[2] The other two that work on international development but have not partnered with J-PAL are the W.K. Kellogg Foundation and the David and Lucile Packard Foundation. The four that we have judged as having a domestic U.S. focus are the Getty Trust, the Robert Wood Johnson Foundation, the Lilly Endowment Inc., and the Andrew W. Mellon Foundation.
[3] When measured by official development assistance granted, including The World Bank, the African Development Bank Group, The Global Fund, the Asian Development Bank, the International Monetary Fund, Unicef, UNRWA, Inter-American Development Bank, the United Nations Development Program, and the World Food Program.
[4] When measured by annual budget. These are World Vision, Save the Children, Catholic Relief Services, CARE, Médecins Sans Frontières, Oxfam, International Rescue Committee, and Mercy Corps.