MIT CEEPR

MIT Center for
Energy and Environmental
Policy Research

Running Randomized Field Experiments for Energy Efficiency Programs: A Practicioner's Guide

Raina Gandhi, Christopher Knittel,
Paula Pedro, and Catherine Wolfram

July 2016 CEEPR WP 2016-015

A Joint Center of the Department of Economics, MIT Energy Initiative and MIT Sloan School of Management.

Running Randomized Field Experiments for Energy Efficiency Programs: A Practitioner's Guide

RAINA GANDHI,[a] CHRISTOPHER R. KNITTEL,[b] PAULA PEDRO,[c] and CATHERINE WOLFRAM[d]

ABSTRACT

Researchers and professional evaluators are increasingly turning to randomized field experiments to evaluate energy efficiency programs and policies. This article provides a brief overview of several experimental methods and discusses their application to energy efficiency programs. We highlight experimental designs, such as randomized encouragement and recruit-and-deny, that are particularly well suited for situations where participation cannot be enforced. The article then discusses several implementation issues that can arise and characterizes applications that are a good fit for a randomized experiment. We also address the most common objections to field experiments, and share the best practices to consider when designing and implementing a field experiment in this space.

Keywords: energy efficiency, evaluation, experiments, randomized trials, best practices

http://dx.doi.org/10.5547/2160-5890.5.2.rgan

1. INTRODUCTION

Field experiments have grown increasingly popular in economics, yet their application to the social sciences overall is relatively new (List, 2009). The purpose of this article is to provide more information about field experiments; describe several methods of conducting them and how to apply them to energy efficiency and conservation interventions ("treatments") of various kinds; and share some findings about how to conduct field experiments within this space.

The practice of conducting randomized experiments is drawn from medicine, where drugs and treatments are tested on animal and human subjects in controlled conditions to assess their efficacy. In economics, randomized field experiments are used to test theories and treatments among humans in a natural, real-world setting (the "field"), where participants face the incentives, constraints, and settings that govern their daily lives. These experiments have grown in popularity because of their unique ability to assess causality and because policy-makers are becoming increasingly focused on "evidence-based" decisions.

Field experiments are particularly interesting and well suited for the energy efficiency and conservation space (Allcott and Greenstone, 2012). Empirical estimates of energy savings and

[a] Product Marketing Analyst at FirstFuel Software.

[b] William Barton Rogers Professor of Energy Economics, Sloan School of Management; Director, Center for Energy and Environmental Policy Research, MIT; Faculty Co-Director, The E2e Project.

[c] Corresponding author. paulagpedro@berkeley.edu Program Manager at The E2e Project.

[d] Cora Jane Flood Professor of Business Administration, Haas School of Business; Faculty Director, Energy Institute at Haas; Faculty Co-Director, The E2e Project.

Economics of Energy & Environmental Policy, Vol. 5, No. 2.

program impacts have primarily come from engineering-style analyses, which do not necessarily reflect real-world conditions (see, e.g., McKinsey 2007), or from observational studies, which cannot always isolate program impact from other factors. Though these methods are useful, they are substantially less rigorous than randomized field experiments because they cannot determine causality, isolate treatment impact or account for unobserved factors—all of which combine to often either overstate or understate results. Furthermore, the need to accurately estimate the impact of energy efficiency and conservation-related policies is particularly important because climate change policies rely substantially on future energy efficiency improvements to generate emissions savings at a low cost. Accurately measuring savings is crucial to ensuring that public policies are achieving their desired goals.

Randomized trials can also provide general insights into how consumers make decisions about energy consumption. Consumers are highly heterogeneous and value products, treatments, and savings differently, suggesting that there are many factors that affect decision-making around energy consumption (Houde, 2014; Allcott and Kessler, 2015; Davis and Metcalf, 2014; among many others). A range of experiments, as summarized in Price (2014) and Hahn and Metcalfe (this issue), has shown that both neo-classical factors (e.g., prices or search costs) and behavioral factors (e.g., salience or social norms) influence energy consumption. A better understanding of this decision-making process is necessary to craft policies that effectively and efficiently achieve their goals.

The article proceeds as follows. Section 2 outlines how field experiments work and several options for when the canonical field experiment, the randomized controlled trial (RCT), does not work. Section 3 describes common objections to field experiments and discusses when these objections are more and less relevant. Section 4 describes the types of programs that are or are not well suited to field experiments and suggests some best (and worst) practices, drawing on lessons from several past field experiments.

2. HOW RANDOMIZED FIELD EXPERIMENTS WORK

Field experiments provide insight into what would have happened to the same participants over the same time period, absent the treatment. In statistical terms, they create credible counterfactuals (Duflo, Glennerster, and Kramer, 2007). Since the true counterfactual can never be observed, it must be approximated by a well-crafted control group. The challenge is to find a valid comparison group: one that is statistically identical to the treatment group and that is equally affected by the same factors as the treatment group (Gertler et al., 2011). However, multiple factors must be considered when choosing a comparison group to prevent bias (Glennerster and Takavarasha, 2013).

To implement a field experiment, the researcher must first recruit potential participants, but this process can introduce selection bias. Consumers who choose to participate in a study may expect to gain the most from participating, perhaps because they believe they will benefit from the treatment. Those who choose to participate, then, are likely systematically different than those who did not. Thus, the estimate of program impact from this group of participants may not generalize to the larger population of interest, which includes those who chose not to enroll in the study. In considering program designs such as this or considering observational analyses, unobserved variables might play an important role: perhaps some households are more environmentally conscious than others, are more willing to participate in efficiency and

conservation programs, and save more energy because of their behavior, overstating the results of the study.

Constructing a sample and comparison group through randomization minimizes these issues. Randomly selecting participants, if done correctly, creates a representative sample, one that mirrors the distribution of characteristics in the population of interest. Randomly assigning participants to the treatment or control group, as in the canonical randomized controlled trial (RCT), ensures that the two groups are statistically identical: that, in expectation, there are no systematic differences between the two groups that could bias the results. Thus, a correctly designed and implemented randomized evaluation provides an unbiased estimate of the impact of the program in the study sample (Gertler et al., 2011). One note here is that the sample has to be of a sufficient size to create a valid comparison group and detect an impact (Imbens and Rubin, 2015), an issue that will be discussed later in this paper.

RCTs identify potential participants for a target population and randomly assign them to either receive the treatment or not. Researchers can then compare outcomes for the control group to those for the treatment group, and, if the RCT is designed and implemented correctly, the difference in outcomes can be attributed to the treatment. For new programs, the treatment group could be a set of households within a utility's service territory that are randomly selected to participate in a pilot. The control group would be households that did not participate. If the evaluation shows that a program produces net energy savings, it can be rolled out to all households.

It may not be possible to assign treatment and control groups. In some cases, participants cannot be mandated to receive a program or treatment, and in other cases, participants cannot be denied access to a program or treatment. For example, a field experiment assessing the impact of home energy audits on energy efficiency investments cannot force people to get audits of their homes. Similarly, customers randomized into a control group cannot be prevented from utilizing rebates that are already offered by the utility.

In other instances, treatment and control groups may be assigned, but compliance to the group assignments may be imperfect. Some participants who were supposed to receive treatment may not get it, or participants in the control group may receive treatment. This can happen if eligibility cutoffs are not strictly enforced, if selective migration takes place based on treatment status, if there are administrative or implementation errors, if some participants in the treatment group choose not to participate, or for many other reasons.

The next section discusses "recruit and deny" (or delay) strategies and "randomized encouragement designs," which can and should be used to account for these factors.

2.1. Recruit and Deny (or Delay) Strategies

Recruit and deny and recruit and delay designs are ideal for situations when participation cannot be mandated or denied, when there is a resource or administrative constraint that limits the number of people who could receive the treatment, and/or when compliance is a concern.

Recruit and deny designs, also known as lottery or oversubscription methods, first have potential participants indicate interest in the program (perhaps by signing up) and then use a lottery to randomly select participants from this group to receive the program. This lottery will create a treatment group (those selected) and a control group (those not selected), and the two groups can be compared to assess the impact of the program *within recruited participants*.

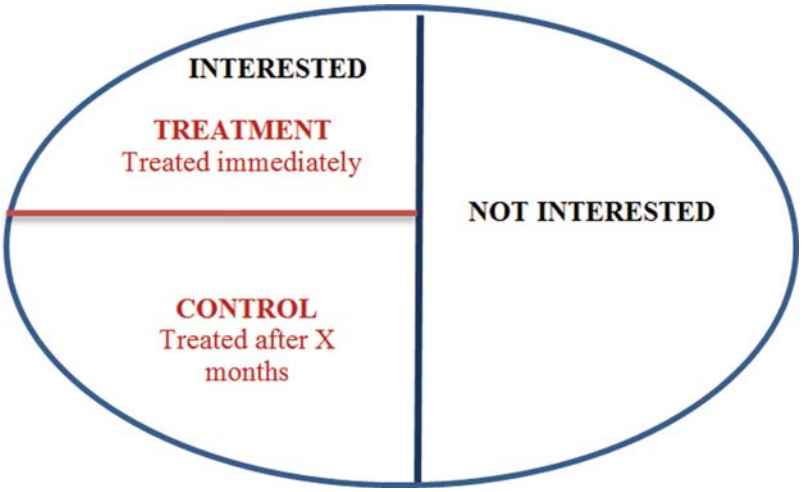

FIGURE 1
Recruit and Delay Design

Randomization can also be introduced by phasing in a program over time, if the order of the phase-in is random. Those who have received the program at a given point in time serve as the treatment group, while those who are still waiting to receive the program are the control group. This version of the design—also known as randomized phase-ins or randomized rollouts—may also incentivize subjects to maintain contact with researchers, avoiding issues with attrition, since the subjects expect to eventually receive the program.

As an example, consider a new energy efficiency program, where the utility or agency implementing the program has limited resources to implement the program across a state and can only cover 10 counties at a time. Instead of comparing the 10 counties that receive a program at the same time with the rest of the counties that do not receive the program or comparing counties before and after they receive the program, researchers could introduce randomization and expand the number of treated counties by doing a randomized phase-in. The counties across the state can be randomly ordered and split into groups of 10 counties each, and the program can be rolled out 10 counties at a time. Then counties that have received the treatment can be compared to those that have not, since the order of the treatment was random, and we can also compare counties before and after they received the treatment.

In order to ensure that the design is valid, researchers need to ensure that cohorts are well identified so that those who have not received the treatment serve as a valid control group. Additionally, the time between phases must be long enough to begin seeing treatment effects.

Though these designs are good ways to introduce randomization into scenarios where it may be difficult, they have some important limitations. Both designs pre-screen the sample for interest, so non-compliance with the treatment assignment likely will not be a problem. However, this means that the treatment is being evaluating in the context of a specific subset of the population: people who would sign up to participate in a program. If the program provides cheap LEDs for home use, for example, households that are interested may be more "green" than those that are not, so that the study sample may not be representative of the target population. What the evaluation assesses is the impact of providing cheap LEDs to homes that are already interested in the program, not on all households. At the same time, if

TABLE 1
Never-taker, always-taker, and complier participation

	When encouraged (treatment)	When not encouraged (control)
Never-takers	Don't enroll	Don't enroll
Always-takers	Enroll	Enroll
Compliers	Enroll	Don't enroll

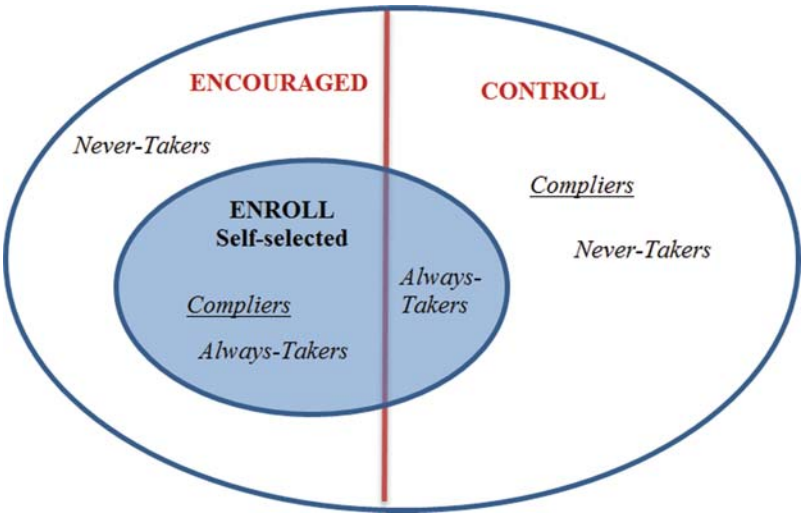

FIGURE 2
Never-taker, always-taker, and complier participation

the program is rolled out more broadly, households that volunteer for the pilot may be very similar to households that take-up the program once offered. However, if carefully constructed, Recruit and Deny (or Delay) designs can provide valuable insight into programs where participation cannot be mandated or denied.

2.2. Randomized Encouragement Design

A second type of strategy that can be adopted in the face of imperfect compliance is the Randomized Encouragement Design (RED). These designs are ideal for situations where randomizing access to a program or mandating participation is neither practical nor desirable. Here participants are randomly selected to be *encouraged* to receive the program (e.g., randomly choosing households to send information about how helpful home audits are). By randomly manipulating the probability that someone receives a treatment, this design can still isolate the treatment's impact.

The theory behind this design is that there are three types of people: people who always enroll in programs (always-takers); people who never enroll (never-takers); and people who participate when encouraged (compliers). The study sample will consist of all three types, as does the overall population, and Table 1 shows how they respond to treatment. The treatment

induces randomly selected compliers to participate in a program, and this difference in program participation and outcomes allows for estimation of program impact.

REDs are also useful in situations where the effects of both participation and outreach/encouragement are of policy interest, and where an encouragement intervention can significantly affect probability of treatment. These designs can also help estimate the cost associated with convincing a customer to do something (usually, buy a product or service), or cost of customer acquisition.

Disadvantages of REDs are that since the treatment cannot be mandated or denied, power calculations must be adjusted to account for lower treatment uptake. The total sample then must be larger than if the same treatment were being evaluated with a canonical RCT. There are additional implementation issues that must be considered, though they will not be discussed here. Duflo, Glennerster, and Kramer (2007) provide a good discussion of these issues.

Note that RCTs with imperfect compliance can be treated as unintentional REDs, as the treatment assignment only influences the probability that someone receives a treatment. Thus, even if participants do not comply with treatment assignments, the program can still be evaluated, though with less statistical precision. For further discussion of REDs, see Gertler et al.'s (2011) discussion of randomized promotion designs.

2.3. Common characteristics of programs that could be used for randomization

Randomization can be introduced into programs through three basic elements: access; timing; and encouragement. We can choose which people are offered access to a program, when people are offered access, and which people are given encouragement to participate—all of which would create an experimental evaluation.

For example, before offering a rebate to all customers, an RCT offering rebates at varying amounts can evaluate how price sensitive customers are and estimate the free-ridership rate. If rebates already exist, an RED can test whether encouragement would increase take-up and how that would affect energy consumption.

Participation cannot be mandated for most programs, so REDs and Recruit and Deny (or Delay) methods can help estimate the impact of the program and provide insight into barriers to program participation. This could be for new programs (to estimate their effectiveness) or existing programs (to estimate effectiveness and understand barriers to adoption).

To pilot a new technology, such as an energy management system, or a program subject to administrative constraints, recruit-and-deny or recruit-and-delay methods can be used to evaluate how effective the technology or program is before it is made widely available. These programs can be pitched as having limited capacity with participants chosen by lottery.

If an RCT was being conducted, but there was imperfect compliance with assignments—whether because some in the treatment group refused treatment or because some in the control group somehow received treatment—the program can still be evaluated, but with the methodology of an RED.

That said, not every program can and should be evaluated with a field experiment. Programs that should be evaluated with a field experiment should meet at least one of these criteria:

Untested. There is little, rigorous evidence about how effective a program is, whether in a particular context or globally. It could be an innovative approach that has never been

applied before or an approach that is widely used but has never been rigorously evaluated.

Affected by behavioral components. For example, providing people with more efficient set-top boxes likely will not induce them to watch more TV, but people with Priuses may drive more since the cost of driving is lower. Evaluations can provide insight into behavioral changes induced by the program, so future programs can be better tailored to achieve maximum impact. That said, almost all programs involve human behavior at some level.

Expensive. Regulators or program managers may decide not to evaluate small-scale or short-term programs using a rigorous field experiment. Also, before committing to spending large sums of money on a large-scale program, it is often useful to conduct a smaller-scale trial or pilot to better understand what effect the full program could have.

Replicable. The program has the potential to be scaled up or implemented in a different context.

Strategically relevant. The program could create significant savings, it is a flagship initiative, or its results could be used to inform key policy decisions.

When considering evaluation design, the three most important criteria are that the design be: (Pritchett 2005)

Technically correct. The study must be internally valid, with indicators that accurately represent the underlying behaviors we are trying to measure. The design should also include protocols to handle potential problems that may arise. It must generate credible.

Politically feasible. The design should be one that would be approved by ethical groups such as university committees for the protection of human subjects.

Administratively implementable. The study design should be made as logistically easy as possible, while still answering the research question. The more difficult the study is to implement, the more likely it is that mistakes will be made that could bias or otherwise interfere with the data and render the results useless.

3. COMMON OBJECTIONS TO FIELD EXPERIMENTS

Below we discuss several commonly raised objections to field experiments. In general, it is important to consider the extent to which objections apply to *any* form of evaluation, or whether they apply to field experiments compared to an alternative evaluation approach. Often the tradeoff is between a less rigorous evaluation approach and a field experiment. In the extreme, regulators are sometimes left assuming that an energy efficiency program performed exactly as expected and use engineering estimates to describe ex post savings. While this may be inexpensive and fast, it amounts to abandoning evaluation, which might be the best approach for small, low-stakes, short-term programs. Ultimately, regulators and program managers must decide how valuable it is to isolate and identify the causal impact of a program.

3.1. They are unfair and unethical

Some object to RCTs on fairness grounds as they see excluding eligible participants as unfair. Randomizing which customer receives a treatment, however, does not necessarily mean denying some consumers the treatment's benefits. Financial and administrative resource con-

straints often do not allow for everyone who would benefit from the program to enroll simultaneously, so randomizing is often the fairest way to allocate treatment order. All eligible beneficiaries have an equal chance of being selected first. This is especially important when participating in programs is highly desirable (Gertler et al., 2011).

Also, as we have noted above, REDs permit everyone who wants to participate to continue doing so, and instead randomize the level of encouragement to participate in the program. An example of this is a famous encouragement design conducted by Sexton and Hebel (1984) in order to assess the impact of smoking on fetal birth weight. In the experiment, researchers encouraged the treatment group not to smoke.

It can also be argued that not conducting an RCT or rigorous impact evaluation is unethical. There should be some data or results to justify investing significant public resources into a program so that public resources are not wasted on an ineffective program. RCT results can also help fine-tune a program to make it more effective and efficient before the program is scaled up or implemented elsewhere.

Most pilot programs are conducted to test the logistics of implementing a program and to gain a preliminary understanding of a program's effects.

Finally, field experiments typically require review and approval from the Institutional Review Board (IRB). The role of the IRB is to ensure that research involving humans is ethical, meets federal, state, and institutional guidelines for the protection of the rights and welfare of human subjects, and minimizes harm to participants.

Any research conducted on human subjects requires approval from the IRB prior to launch, where research and human subjects are defined as follows:

> *Research* means a *systematic investigation*, including research development, testing and evaluation, *designed* to develop or contribute to *generalizable knowledge*.
>
> *Human Subject* means a living individual *about whom* an investigator conducting *research* obtains: (1) data through *intervention* or *interaction* with the individual or (2) identifiable *private information*.[1]

As an example, a field experiment considering the impact of energy technologies on commercial and industrial customers' energy consumption may not require IRB approval, since commercial and industrial utility customers do not qualify as human subjects. However, if the study required surveys that collected information about individuals working at those firms, it may qualify as research on human subjects.

Another example is that reviews of preexisting data may not require IRB approval, if the data include no identifiable private information.

There are some exceptions to these definitions, so the determination that the research project does not need IRB approval should be made by the IRB rather than the researcher.

All personnel involved with conducting research with human subjects must have completed the Collaborative IRB Training Initiative (CITI) Human Subjects Protection Training, and this must be updated every three years. Four to six weeks should be allocated for IRB submission and the review process, and all approval or determinations that a project does not require IRB review should be saved.

1. Though these definitions are federally defined in 45 CFR 46.102, the emphasis here is added by UC Berkeley's Committee for the Protection of Human Subjects, available here: http://cphs.berkeley.edu/review.html.

3.2. They are too expensive

Multiple resources are necessary to implement an experiment, and these all have costs. The evaluation needs to be built and implemented upfront. Costs include creating the program materials, creating and pilot testing surveys, data collection materials, training for staff, staff wages, data entry operations, and more. These costs are sometimes spread over several years. Gertler et al. (2011) provides cost data for several impact evaluations of World Bank-supported projects, though these are not energy-related.

The largest cost is new data collection. In applications to energy consumption behavior, however, most of the data are already being collected, since utilities have data on consumers' energy usage and spending. This drastically lowers the cost of conducting field experiments in this sector when the utilities are willing to or required to make these data available for the RCT.

Additionally, conducting a field experiment for a potential new program or product can provide a much better understanding of a project's cost-effectiveness. Implementing the project on a smaller scale provides a more detailed understanding of the costs of the project and randomizing provides a much more accurate estimate of its benefits, so results from a field experiment would be very useful for decision-makers deciding to scale up or replicate a program (List, 2011).

Finally, embarking on any research project has costs. Randomized field experiments do not have to be more expensive than other evaluations; they are, if anything, less expensive than most methods, since they rely on the study design to draw inferences about behavior instead of relying on surveying people after a treatment.

3.3. They take too long to run

Experiments need to be timed appropriately. Field experiments of programs must be conducted at the same time as the pilot or the rollout, not afterwards; otherwise, the evaluation will not be able to demonstrate a direct program impact. However, they should be implemented once the program is fairly well established; if the content or logistics are likely to be revised significantly, an evaluation would likely have less impact on future decisions.

The experiment should match the program implementation cycle, continuing for at least as long as the program. It also needs to allow enough time for the treatment to have an effect. For example, an evaluation of an informational campaign on energy efficiency investments that ended a week after the treatment could drastically underestimate the effect of the program, since most people and businesses do not make these investment decisions that quickly. People generally take time to learn and adapt their behavior, and so indicators should not be measured too early. Instead, follow-ups should be conducted for an appropriate amount of time, depending on the specific program and what is logical.

3.4. They can only be done at the individual level

In some circumstances, randomizing at the individual level is not possible or not desired. An alternative is clustered randomization, where groups of individuals, such as schools or neighborhoods, are randomized.

For example, Cornelius et al. (2014) evaluate a school-based intervention to promote energy- and GHG-saving behaviors. In their intervention, treatment was composed of a five-week curriculum promoting changes to reduce energy use and GHG emissions (home elec-

tricity, transportation, and food-related). The experiment was clustered within classrooms. The advantages of such an approach are that it can provide insight into spillovers (e.g., across classrooms within a school), can make compliance with treatment assignment and implementation easier logistically, and can reduce perceived unfairness.

A clustered approach tends to require, however, a larger sample size to maintain power. Duflo, Glennerster, and Kramer (2007) provide a more detailed analysis of the advantages and disadvantages of randomizing at different levels.

3.5. They can only be performed on the residential sector

Field experiments can be performed on individuals and individual households, but can also be applied to the commercial and industrial sectors. For example, businesses can be induced to participate in energy efficiency programs through REDs, or large industrial customers can be recruited to pilot a technology as part of a randomized phase-in. Much of the current energy efficiency research has focused on the residential sector, but the commercial and industrial sectors represent large savings potential and field experiments are equally valid.

4. OPERATIONAL DOS AND DON'TS

4.1. Develop the program theory or logic model

Carefully thinking about the logic model behind the experiment enables better planning and better field implementation. This means meticulously mapping the question being asked, the outcomes of interest, and the goals of the program. This exercise not only provides a better understanding of how the program is expected to affect participants, but also enables the researcher to better design an experiment that tests the exact mechanisms through which the program is expected to work.

Consider the following problem: electricity consumption in California's households spikes during extreme heat events, leading to dispatch of high cost generating units and increased risk of outages. A utility company would like to test whether smart residential meters and in-home displays for households (input) might help curb peak consumption in its territory. The strategic objective is that households would become more aware of their electricity consumption patterns with those devices. The output of the program is the delivery and installation of the input (smart meters and in-home displays). The evaluation will first look at whether households "use"/interact with the in-home displays (intermediate result). The final result (or goal, or outcome of interest) is whether households decrease their electricity consumption overall and during peak time.

For any intervention or program evaluation, the treatment must be well defined (Holland, 1986). Consider, for instance, two conclusions: the first one is "she saved energy because she is young and tech-savvy" and the second one is "she saved energy because she received a smart-meter." The treatment in the first scenario (being young and tech-savvy) cannot be manipulated and thus is not well defined. On the other hand, the second scenario has a well-defined treatment and therefore can have its effect isolated on a field experiment.

Articulating the logic model behind the treatment and testing theories tends to lead to more generalizable results by allowing us to better understand the mechanisms through which a treatment drives change.

4.2. Discuss the validity of the experiment (internal, external, construct)

While the field experiment is being designed, issues of validity (internal validity, external validity, construct validity, and economic validity) should be considered.

Internal validity relates to the ability to draw casual inferences from the data. In other words, can we attribute the differences observed between control and treatment to the program itself, or is something else causing the change? This is perhaps the most important aspect—almost serving as a pre-requisite—for a field experiment. Absent random assignment, there is the risk that systematic differences might be responsible for some of the observed differences in the outcome of interest. To ensure that the evaluation is internally valid, the experiment needs to adhere to a few assumptions, described below.

External validity describes the extent to which a study's results can be generalized or applied to other subjects or settings. Studies implemented in computer labs, for instance, are often subject to criticism regarding external validity because conditions given to participants (and the participants themselves) are often too different from real-life situations. The same happens with studies transposed from one country to another, where conditions and institutions are too different. For example, are the results of an evaluation conducted with residential homeowners in Massachusetts in the 1990s applicable to commercial & industrial energy users in Mumbai in 2014? Probably not. Is this same study applicable to residential homeowners in California? Maybe.

To establish the external validity of a finding, ideally researchers would replicate experiments in different settings and different populations. That said, in the context of one specific study, very rarely can (or should) the treatment be administered to a truly representative sample of the whole population.

We recommend that researchers consider the following issues before deciding on the study's sample:

Who is this intervention meant to serve, and what is the larger population of interest?
Does the chosen sample allow the experiment to answer the proposed questions?
What can be done given the available resources (e.g., financial, political)?
How does this study fit within the literature? Does it corroborate the external validity of another study?
How well does the design allow us to evaluate the logic model?

Construct validity refers to the degree to which the experiment measures what it claims to measure. Unless the field experiment is conducted in a naturalistic manner, some feature that is unique to the experiment design might generate results that are idiosyncratic or misleading.

For instance, asking clients how much energy they use has less construct validity than directly looking at their usage bills, since they likely would not remember how long they left the lights on in each room. Another example of low construct validity in a study would be if the researcher decides to pay the team in charge of delivering the program per its performance, when this will not be the case after the program gets scaled up—likely yielding significant differences in take-up and results.

Economic validity relates to how large the observed effect is, and whether that will have practical significance. A program may have a statistically significant effect, but if the effect is very small, it may not justify the costs and effort of scaling up that program.

4.3. Avoid situations where the set-up of the experiment biases participants' behavior

Participants might change their behavior when aware of their participation in the evaluation and/or their treatment status. These biases might distort experimental results and researchers should try to anticipate (and mitigate) them as much as possible:

The Hawthorne effect happens when individuals change their behavior because they are aware of being observed. For instance, if participants know that their energy consumption data will be shared with researchers, they might pay more attention to their habits and save energy that they wouldn't have saved absent the study.

The John Henry effect can affect participants in the control group. These participants might feel offended or resentful for not receiving the treatment and might either work harder to compensate for the lack of treatment or slack off in their behavior.

The Placebo effect can also affect participants in the control group. For example, if a study involves installing new smart meters to collect data from both the treatment and control groups and customers begin accessing more detailed information about their usage, the control group may change their behavior as a result of the study.

The most effective way to protect a trial against these biases is by keeping participants unaware of the intervention and of the assignments for as long as possible. In medicine, this is called "blinding" and/or "masking." In economics, this can be harder to operationalize. Utility companies (who often govern access to billing data), partner organizations, and universities' institutional review boards, which oversee research involving human subjects, often require that participants sign consent forms, thus making them aware of the trial. Depending on how risky the trial is deemed by the institutional review board, researchers may be able to request waivers if they can explain that the Hawthorne could significantly bias results. Likewise, they might be able to negotiate this aspect with the IOUs and partner organizations, depending on their own bylaws. It is also preferable to keep the treatment status confidential (to the participants and to the program administrators) whenever possible.

4.4. Perform statistical power calculations

One of the first steps a researcher should take before designing an experiment is to calculate the sample size necessary to be able to identify a treatment effect of a particular size---a so-called power calculation. Well-conducted power calculations ensure that the analysis correctly detects the effect of the program while minimizing the data and resources involved. An underpowered study does not involve enough participants, leading to inconclusive results and wasting the time and the money invested on data collection. An overpowered study involves more than the required number of participants, also wasting valuable resources that could be used for another purpose.

The canonical power equation maps the relationship between the minimum effect that can be detected by the analysis for a given control and treatment group sizes. This is also called the minimum detectable effect (MDE). As an example, suppose that the power equation results are such that for control and treatment groups of 1,000 customers each, the MDE is 3%. Additionally, suppose that the real impact of the treatment is of 2%. Given that 3% is greater 2%, it would be impossible to (statistically) detect the impact of the program for this size of control and treatment.

The more precise the test or the smaller the treatment effect, the harder it is to infer causality with the data. In terms of precision, typically the industry standard is of 80% power

and 5% significance, meaning that there is an 80% chance of correctly detecting a difference between treatment and control when there is a difference and 5% chance of detecting a difference between treatment and control when there is no real difference.

The precision of the test also depends on behavioral patterns. First, variance is a measure of how much the behavior of participants in the sample changes over time absent the treatment. Intuitively, the less it varies, the easier it is to detect the program's impact. This is because changes in energy usage can be attributed to the program, not to chance or natural variation. This means that lower variance allows for a smaller sample size. The second factor is the intra-cluster correlation, which is a measure of the correlation in behavior across participants in the study. The idea is that participants might share similarities that make them react similarly when exposed to a common shock. The higher the intra-cluster correlation, the harder it becomes to distinguish the impact of the intervention from another shock to participants (and the larger the sample needed). These patterns can be estimated using baseline data, data from a similar survey, or the most similar data available to the researchers.

If the field experiment is a randomized encouragement design, researchers also need to estimate the expected take-up rate. This is usually done by either talking to the team in charge of the implementation or by talking to staff members at the partner organization.

Finally, if it is significantly more expensive to have customers in the control or the treatment group, the proportion of customers assigned to each group can be adjusted, thereby providing more precision to the experiment. Researchers should strive to minimize the MDE subject to the evaluation budget, and this is an area that can be explored with statistics.

For more information on conducting power calculations in randomized trials, please refer to Duflo, Glennerster, and Kramer (2007) and Gelman and Hill (2006).

Although quite useful, it is important to highlight that the results of power calculations should be taken as a suggestion for the approximate size of the sample, not of the exact sample size. Power calculations rely on forecasts, and these can turn out to be wrong, so it makes sense to treat them as indicative but not dispositive.

4.5. Develop and file a pre-analysis plan

A pre-analysis plan is a document that describes the program framework, the evaluation approach, potential threats to external or internal validity, power calculations, the data collection and manipulation, and the equations to be estimated after the intervention is completed.

Pre-analysis plans are required in medicine[2] and have become increasingly popular in economics because of the protection that they offer to the researcher from both the criticism and the temptation to data-mine and cherry-pick (Casey, Glennerster, and Miguel 2012). These plans also help the researcher think through the implementation and the analysis of the study before the launch. They provide a roadmap for the analysis, making the delivery of the results—once the intervention is over—speedy and relevant.

These documents can be made public at the Social Science Registry's website.[3]

2. The International Committee of Medical Journal Editors (ICME), "requires, and recommends that all medical journal editors require, registration of clinical trials in a public trials registry at or before the time of first patient enrollment as a condition of consideration for publication." Source available here.

3. www.socialscienceregistry.org

4.6. Treat control and treatment groups *exactly* the same

During the experiment, all customers should be handled in exactly the same way, with the obvious exception that the treated group receives the treatment. Although intuitive when thinking about the design of the experiment, this issue might be challenging to guarantee in the field, especially when experiments involve several different treatment arms.

As an example, it is often the case that the partner or government organization has prepared promotional materials (TV commercials, brochures) about the program that cannot be restricted to just the treatment group. If that is the case, it is crucial to ensure that both groups are exposed equally, on average, to the same materials. This also means that the effect of this exposure will not be captured by the evaluation, which might underestimate the impact of the program.

Another type of bias could arise when the organization knows which group the participant was assigned to prior to the delivery of the program. If a participant is in a treatment group that receives a higher financial incentive, for instance, the organization might be tempted to put more effort into enrolling this participant, creating differences between the treatment and control groups in addition to the financial incentive. We recommend that, in cases where incentives might be misaligned, the assignment is only revealed later on in the process.

Consider an evaluation assessing the impact of providing households with information about their energy efficiency during an audit. If this evaluation relies on surveys administered at the end of the audit and the partner is aware of the treatment assignment, they may believe that it is more important to induce the treatment group to complete the survey rather than the control group. The treatment group then is treated differently than the control group and is likely to have a higher survey response rate. If the response rates between the treatment and control are not comparable, the internal validity of the study is compromised.

It is also just as important to ensure the treatment is relevant, when compared to everything else that both the control and treatment would be exposed to. Suppose an organization wants to understand the impact of an energy efficiency program on small businesses. This program is composed of a series of marketing campaigns that entice participants to commit to saving energy. With the intention of evaluating this program, a random encouragement design is implemented and a randomly selected group of these small businesses receives a letter inviting them to participate in the program. In theory, as long as all businesses in both groups are exposed to same materials by the same amount, the randomization is still valid. However, the bigger the overall marketing campaign, the more difficult it is to tease out the impact of one particular tool.

4.7. Adhere to assumptions that guarantee identification

In order to identify the causal impact of an intervention, the experiment has to adhere to a set of assumptions (Imbens and Rubin, 2015)[4]:

Unconfoundedness: Briefly, this assumption requires that a customer's assignment to either the treatment or control group is not a function of their expected reaction to the treatment. In the case of an experiment, this is guaranteed because of the random assignment—if the assignments are truly random. Unfortunately, there are no good ways to prove

4. There are several other technical assumptions, including "individualistic assignment," which ensures that a customer's assignment is not a function of other customer's covariates or potential outcomes, and "probabilistic assignment," which ensure that no customer is guaranteed to be in either the treatment or control group. See Imbens and Rubin (2015) for more detail.

this. We recommend, however, that researchers make the software codes used to define the random assignments available to the public. It is also conventional to demonstrate that variables measured at baseline have the same means across the groups as an indication of balance, although this should not be seen as "proving" balance as it only holds for the variables that can be measured.

Stable unit treatment value assumption (SUTVA): This assumption requires that whether a subject is exposed to the treatment depends only on the subject's own assignment, not on the assignment of other subjects. Second, in the case of a RED, the subject's outcome is a function of her encouragement status and treatment status, but not a function of the encouragement or treatment status of other subjects. This assumption might be violated if neighbors discuss their program participation with each other, for instance.

For a Randomized Encouragement Design, two more assumptions are needed:

Monotonicity: The encouragement must positively affect program participation for all encouraged participants. This means that the encouragement can never decrease program participation, although it can be the case that the encouragement has no effect. More generally, monotonicity implies that there are no "defiers" in the sample. Here, too, there are no good ways to demonstrate that this assumption holds in an experiment. A test that could provide an indication is to compare commitment rates in the encouraged and control groups and demonstrate that this rate is significantly higher in the encouraged group.

Exclusion restriction: The encouragement cannot directly affect the final outcome of interest. For the conclusions of the experiment to hold, we need to assume that the encouragement only affects the outcome indirectly, via changing program participation. If the encouragement gets participants thinking—and acting—differently, this could introduce bias into the estimates.

Suppose we have designed a RED to understand the impact of a certain program on energy consumption. If, for instance, the encouragement reveals information about your energy consumption, encouraged consumers might change their energy consumption because of the encouragement and not because they have enrolled in the program. Estimates of program impact would then incorrectly attribute the observed change in energy usage to the program, when some of the effect came from the encouragement.

4.8. Monitor the implementation

It is crucial to monitor that the intervention is being adequately implemented to the treated participants and that the control group is not being contaminated (receiving the intervention through some other means).

Contamination occurs when a participant is given a treatment other than the one that she was originally assigned in the study. To minimize mistakes and identify sources of systematic contamination, we recommend that researchers and partner share a weekly/bi-weekly report with assignment and location/field workers.

Spillovers occur when the treatment also affects the outcome of the control group. If, for instance, the treatment includes tips on how to save energy given to households in a very urbanized neighborhood, we recommend that the researchers change the level of the randomization. Instead of at the household level, randomize at the neighborhood level.

Pioneer or partial equilibrium effect early effects of an RCT may be quite different from later effects, and sometimes yield impacts in opposite directions. This is because learning

effects can take some time. This means that short-term experiments should be interpreted with caution.

4.9. Be transparent when reporting results

As mentioned above, the study should contain a table that compares groups, on average, for a series of observable variables, testing for statistical differences between them. In the case of energy efficiency, this might include size of the household/business, zip code, NAICs (for commercial and industrial), average monthly/daily energy usage, etc. This does not guarantee identical groups (on observables and non-observables) but gives a sense of balance.

The reporting should also follow the pre-analysis plan as closely as possible and estimate all the equations specified in the document. Any departures from the pre-analysis plan should be highlighted.

Finally, we recommend that researchers think about cost-effectiveness calculations prior to the launching of the study. This guarantees that relevant variables can be incorporated into the data collection. For instance, energy efficiency programs typically claim to have other, non-energy benefits such as providing more comfortable, less drafty houses. If that is the case, in-house temperature could be collected so that this benefit can also be taken into consideration in the evaluation of the program.

5. CONCLUSION

Field experiments are not always applicable or the best method for a given research question. Yet field experiments to evaluate consumer and firm energy efficiency behavior and interventions designed to affect energy consumption behavior can provide useful insight into how people consume energy and make decisions about energy efficiency investments that can broaden our understanding about the most effective programs and policies.

Useful field experiments should be unbiased and rigorous, based on the best methodology available and implemented correctly. They should be substantive, providing novel insights and focusing on areas that lack definitive evidence. Replicating field experiments is also critical, providing insight into how generalizable the results are across settings and contexts. They should also be relevant, timely, and actionable to be useful both for academics and for policymakers and practitioners.

REFERENCES

Allcott, Hunt and Michael Greenstone (2012) "*Is There an Energy Efficiency Gap*?" Journal of Economic Perspectives, No. 26 Vol. 1.

Allcott, Hunt and Judd Kessler (2015) "*The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons*" NBER Working Paper 21671.

Bloom, Howard (2005) "*Randomizing Groups to Evaluate Place-Based Programs*" Russell Sage Foundation, chapter "Learning more from social experiments".

Casey, Katherine, Rachel Glennerster, and Edward Miguel (2012) "*Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan*" Quarterly Journal of Economics, No. 127 Vol. 4.

Cornelius, Marilyn, Carrie Armel, Kathryn Hoffman, Lindsay Allen, Susan Bryson, Manisha Desai, Thomas Robinson (2014) "*Increasing Energy- and Greenhouse Gas-Saving Behaviors among Adolescents: a School-Based Cluster-Randomized Controlled Trial"* Energy Efficiency, Vol. 7 Issue 2.

Cox, David and Nancy Reid (2000) "*The Theory of the Design of Experiments*" Chapman & Hall/CRC.

Davis, Lucas and Gilbert Metcalf (2014) "*Does Better Information Lead to Better Choices? Evidence from Energy-Efficiency Labels*" NBER Working Paper No. 20720.

Duflo, Esther, Rachel Glennerster, and Michael Kremer (2007) "*Using Randomization in Development Economics Research: A Toolkit*" Centre for Economic Policy Research Discussion Paper 6059.

Gelman, Andrew and Jennifer Hill (2007) "*Data Analysis Using Regression and Multilevel/Hierarchical Models"* Cambridge University Press.

Gertler, Paul, Sebastian Martinez, Patrick Premand, Laura Rawlings, and Christel Vermeersch (2011) "*Impact Evaluation in Practice"* World Bank Publications.

Glennerster, Rachel and Kudzai Takavarasha (2013) "*Running Randomized Evaluations: A Practical Guide*" Princeton University Press.

Hahn, Robert and Robert Metcalfe (2016) "*The Impact of Behavioral Science Experiments on Energy Policy*" Economics of Energy and Environmental Policy, this issue.

Houde, Sebastien (2014) "*How Consumers Respond to Environmental Certification and the Value of Energy Information*" NBER Working Paper No. 20019.

Imbens, Guido and Donald Rubin (2015) "*Causal Inference for Statistics, Social, and Biomedical Sciences—An Introduction*" Cambridge University Press.

List, John and Robert Metcalfe (2015) "*Field Experiments in the Developed World: An Introduction*" Oxford Review of Economic Policy, No. 30 Vol. 4.

List, John (2009) "*Introduction to Field Experiments in Economics*" Journal of Economic Behavior and Organization, No. 70 Vol. 3.

List, John (2011) "*Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off*" Journal of Economic Perspectives, No. 25 Vol. 3.

McKinsey & Company (2007) "*Reducing U.S. Greenhouse Gas Emissions: How Much at What Cost?"* Available here: http://www.mckinsey.com/business-functions/sustainability-and-resource-productivity/our-insights/reducing-us-greenhouse-gas-emissions.

Price, Michael (2014) "*Using Field Experiments to Address Environmental Externalities and Resource Scarcity: Major Lessons Learned and New Directions for Future Research*" Oxford Review of Economic Policy, No. 30 Vol. 4.

Pritchett, Lant (2005) "A Lecture on the Political Economy of Targeted Safety Nets" The World Bank's Social Protection Discussion Paper Series no. 0501.

Sexton, Mary and Richard HEBEL (1984) "*A Clinical Trial of Change in Maternal Smoking and its Effect on Birth Weight"* Journal of the American Medical Association, No. 251 Vol. 7.

APPENDIX—ADVANTAGES AND DISADVANTAGES OF AVAILABLE METHODOLOGIES

TABLE 2
Summary of Available Methodologies

	Description	Comparison Group	Advantages	Disadvantages
Ex-ante studies	Studies done before a treatment is implemented, such as engineering studies	None	Can provide valuable insight into the potential of a program	Relies heavily on assumptions Does not mirror real-life conditions
Qualitative impact evaluations	Talking to the participants through direct observation, open ended interviews, closed-ended interviews, focus groups, etc.	In general, they do not attempt to draw conclusions. But when they do, the comparison is what would have happened to the participant absent the program	Richness of the information collected: instead of measuring the energy usage; this would capture the enthusiasm of a family over Opower reports, for example	Needs to be translated into numbers in order to know if statistically significant. Need to trust that the data was correctly summarized Recall bias—people are forgetful People are terrible at estimating their own behavior in an alternate version of the universe Low response rates Sample selection issues Incentives to answer truthfully are often misaligned
Simple before-and-after	Measures how participants improved over time	The comparison is the participant himself, before the program was given to him/her The same data needs to be collected before and after the program	Simple and easy	Assumes that the participant's outcome would have been the same as before, had they not received the program—or that the program is the only factor influencing the outcome over time The participant himself can go through changes that are not caused by the program Does not account for "rebound"—the fact that individuals adjust to conditions

(continued)

TABLE 2
Summary of Available Methodologies *(continued)*

	Description	Comparison Group	Advantages	Disadvantages
Multivariate regression	Similar as before, but "controlling" for other factors that might explain the differences, trying to isolate the impact of the program	Individuals who didn't receive the program	Overcomes some of the selection bias problem	Assumes that you are really controlling for all of the relevant variables Some differences are unobservable, such as ambition
Statistical Matching	Similar to participant-non participant comparison, where you match participants to non-participants who look like them	*Exact matching*: each participant has a match *Propensity score matching*: every individual is given a probability to participate. Each participant is match to someone with the same score	Better version of the regression	Can be difficult to find the right matches Assumes that researchers are correct about which variables are the most important and that there are no confounding variables
Difference-in-Differences	Compares groups over two time periods: before either group receives the treatment, and after one group receives the treatment	Individuals who didn't receive the program	Easy to do Uses observational data	Assumes that the change in outcome for the comparison group is a credible counterfactual for the treatment group
Regression Discontinuity	When a threshold determines who receives treatment, RDDs compare those just above and just below the threshold	Individuals who didn't receive the program (because they were just below the threshold)	Thresholds are ubiquitous in energy, so it has high potential Highly credible near the discontinuity	Estimates are only valid at the discontinuity Can only be applied in certain settings
Natural Experiments	When subjects receive or don't receive a treatment through a process that resembles random assignment	Individuals who didn't receive the program	Useful when controlled experimentation is difficult or unethical More reflective of real life	No control over other variables
Recruit and Delay/ Recruit and Deny	Either randomly choose participants who have indicated interest in a program (e.g. through a lottery) or randomizing the order that a program is phased in	The individuals who had not yet received the program	Introduces randomization when you cannot deny or mandate participation Non-compliance with treatment will not be a problem Lessens control group attrition Useful for when resources are limited	Sample may not be representative of the entire population, because it has already been pre-screened for interest Estimates may not be generalizable Knowing that they will eventually receive the program may change the control group's behavior

(continued)

TABLE 2
Summary of Available Methodologies *(continued)*

	Description	Comparison Group	Advantages	Disadvantages
REDs	Randomly assign participants to be encouraged to receive a program	Individuals who were not encouraged to participate in the program	Introduces randomization when you cannot deny participation Also measures the cost of acquisition Particularly useful when effects of participation and outreach are of policy interest	Low take-up means that you need a much larger sample Can be very expensive
RCTs	Randomly assign participants to receive a treatment or be in the control group	Individuals who didn't receive the program	Randomizing minimizes selection bias and ensures the groups are statistically identical Yield unbiased and consistent results	Results may not be generalizable to all contexts Expensive Must be well-designed and well-executed to be credible, otherwise will produce biased results Issues with compliance and attrition

Source: Elaborated using Glennerster and Takavarasha (2013).