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

1 Motivations

Economics researchers and professional evaluators are increasingly turning to randomized field experiments to measure the impact of energy efficiency programs and policies. Drawn from the field of medicine, these experiments can be used to test theories and treatments in a natural, real-world setting. In the energy efficiency and conservation space, experiments can be particularly interesting and well suited for evaluations. Not only do utility companies already collect energy usage data - significantly lowering the cost of an experiment – but also because climate change policies rely substantially on future energy efficiency improvements. Accurately measuring savings is crucial to ensuring that public policies are achieving their desired goals. This article provides a brief overview of several experimental methods available and discusses their application to energy efficiency programs.

2 Available Methodologies

A randomized experiment produces a credible comparison group that minimizes biases created by unobservable variables. If designed and implemented correctly, the canonical **randomized controlled trial (RCT)** can attribute the difference between treated and controlled outcomes to the program. One defining aspect of the RCT is that participation in the treatment has to be denied/mandated. Unfortunately, this is often impossible or not desired.

Recruit-and-deny (or delay) designs can be appropriate in these cases. Potential participants indicate interest in the program and a lottery is used to randomly select program recipients amongst those interested. Although more flexible, this design only evaluates the impact of the program amongst people who are already interested in the program, who can be systematically different from the general population. A second type of strategy is the **Randomized Encouragement Design (RED)**, where participants are randomly selected to be encouraged to receive the program. These designs are useful in situations where the effects of both participation and outreach/encouragement are of policy interest. The main disadvantage of REDs is that the necessary sample size is larger than if an RCT were employed to evaluate the same treatment. Additionally, the RED analysis focuses on participants who would not have enrolled absent the encouragement ("compliers"), and these participants can also be systematically different from the general population.

Not all programs can or should be evaluated with a field experiment. Researchers should focus their evaluation efforts on programs that are **untested** (or have little rigorous evidence available), **affected by behavioral components, expensive (**to avoid committing large sums of money to a large-scale program in the absence of evidence), **replicable**, and **strategically relevant**. Additionally, evaluators should not only focus on **technically correct** studies. Being **politically feasible** and **administratively implementable** is just as important.

Some object to RCTs as unfair or unethical. Financial and administrative resources, however, often prevent everyone from enrolling in a program simultaneously, so randomizing is often the fairest way to allocate resources, since all eligible beneficiaries have an equal chance of being selected first. RCT results also help fine-tune a program to make it more effective and efficient before the program is scaled up. A second common criticism is that RCTs are too expensive. It is true that any research has costs, but the largest of them tends to be data collection. In the energy efficiency space, the data are already collected, lowering the cost of experiments. A third objection is that they take too long to run. Making sure that the experiment matches the program implementation cycle and committing to deliver results promptly after the data is delivered is key to ensure the policy relevance of the experiment.

3 Policy implications

Throughout the study cycle, several factors should be designed to minimize the risks of biases.

Design. It is important to develop a program theory, or logic model, to better understand how the program is expected to affect participants. The validity of the experiment should also be discussed (e.g. internal, external, construct, and economic) to ensure that the study is not only technically correct but also relevant. The chosen sample size (from power calculations) should also ensure that the analysis will be able to detect the effect of the program while minimizing the data and resources involved. Finally, the design should adhere to assumptions that guarantee identification: unconfoundedness, stable unit treatment value assumption (SUTVA), monotonicity, and exclusion restriction.

Implementation. It is fairly common for the set-up of the experiment to affect participants' behavior, distorting results (e.g. Hawthorne effect, John Henry effect, Placebo effect). These situations should be anticipated and mitigated as much as possible. During the implementation period, it is extremely important that all customers 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. Groups should also be inspected for contamination (when control participants receive the intervention) and for spillovers (when the treatment also affects the outcome of the control group) that can distort the interpretation of results.

Analysis. A pre-analysis plan[1] should be developed and filed before data are transferred to the research team. This helps evaluators protect themselves from both the criticism and the temptation to data-mine and cherry-pick. The pre-analysis plan should document the experimental design and implementation, as well as the equations to be estimated after the intervention is completed and cost-effectiveness analysis. Final results should be as transparent as possible and include a table that compares groups, on average, for a series of observable variables, testing for statistical differences between them (e.g. size of the household/business, zip+4, NAICs, average monthly/daily energy usage, etc.) The reporting should also follow the pre-analysis plan as closely as possible. Any departures from the pre-analysis plan should be highlighted and explained.

1 www.socialscienceregistry.org

4 Conclusions

Field experiments are not always applicable or the best method for a given research question. Yet, in many cases they 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 can provide novel insights and help focus scarce resources on the most cost-effective programs.