Experimentation and the Evaluation of Energy Efficiency Programs: Will the Twain Meet?

Edward Vine, Lawrence Berkeley National Laboratory and California Institute for Energy and Environment
Michael Sullivan, Freeman, Sullivan & Company
Loren Lutzenhiser, Portland State University
Carl Blumstein, California Institute for Energy and Environment
Bill Miller, SRA International

ABSTRACT

The use of experimentation – particularly, randomized controlled trials (RCTs) where subjects are randomly assigned to treatment and control conditions – has rarely been applied to rigorously test alternative energy efficiency program design features and, more fundamentally, determine the benefits of energy efficiency activities. This absence of a sound empirical foundation for calculating energy efficiency impacts is impeding progress in the development of effective energy efficiency programs. Accordingly, this paper discusses the use of experimentation in the energy efficiency program field with the hope that more experimentation occurs. First, a brief overview of experimental methods is presented. This discussion describes the advantages and disadvantages of conducting experimentation in the context of the development and evaluation of energy efficiency programs. It then discusses barriers to the use of experimental methods (including cost and equity issues), and suggests some ways of overcoming these barriers. Finally, recommendations are made for implementing key social experiments, discussing the types of energy efficiency programs and issues that can make use of experimentation and variables that one might use for selecting treatments.

Introduction

"Oh, East is East, and West is West, and never the twain shall meet"[1]

The evaluation, measurement and verification (EM&V) of energy efficiency programs (publicly funded and ratepayer funded) has a rich and extensive history in the United States, dating back to the late 1970s. While the engineering calculations required to assess the impacts of replacing inefficient technology with technology that is more efficient are straightforward and relatively uncontroversial, there are certain critical issues that have resisted resolution and create a persistent climate of uncertainty about the effectiveness of energy efficiency programs. These issues are generally described under the headings of spillover effects (i.e., program impacts that indirectly cause changes in energy use of parties who were not directly treated by the program) and attribution problems (i.e., net-to-gross or changes in energy use that would have occurred in the normal course of history without the intervention). Both of these issues have at their core uncertainty about the counterfactual — that is, what would have happened without the program? We need the counterfactual because, assuming we know what has happened, the

[1] Taken from "The Ballad of East and West," a poem by Rudyard Kipling and first published in 1889.

measure of a program's success is the difference between what has happened and what would have happened without the program. Much of the uncertainty about the counterfactual turns on uncertainty about how human behavior will affect the eventual magnitude of the energy efficiency improvement obtained from program initiatives. Like most other questions about the impacts of program initiatives on human behavior, these issues can really only be resolved through careful scientific study using experimentation.

Unfortunately, the use of experimentation, particularly randomized controlled trials (RCTs), has rarely been applied to rigorously test alternative energy efficiency program design features and, more fundamentally, determine the benefits of energy efficiency policy initiatives. The resulting absence of a sound empirical foundation for calculating energy efficiency impacts is impeding progress in the development of effective energy efficiency programs and has led some in the policy community to advocate non-energy efficiency options with more rigorous foundations and less risk of failure in moving the U.S. along a path toward an environmentally more robust energy system.

Historically, the energy efficiency evaluation community has relied primarily on quasi-experimental methods (e.g., comparing energy impacts for participants and non-participants) for evaluating the impacts and performance of energy efficiency programs. The drawbacks of this approach are well understood in the policy literature, and evaluators are now exploring the use of RCTs and determining their viability in the evaluation of energy efficiency programs. Still, this emerging interest in RCTs rarely extends to the basic question: how much, if anything, does energy efficiency save?

Based on a review of the literature and discussions with program managers and evaluators, this paper discusses the use of experimentation in the energy efficiency program field. First, a brief overview of experimental methods is presented. This discussion describes the advantages and disadvantages of conducting experimentation in the context of the development and evaluation of energy efficiency programs. It then discusses barriers to the use of experimental methods and suggests some ways of overcoming these barriers. Finally, recommendations are made for implementing key social experiments, discussing the types of energy efficiency programs and issues that can make use of experimentation and variables that one might use for selecting treatments.

The use of RCTs and quasi-experimental methods offers an opportunity for program managers to develop more innovative and effective programs. Methodological tools are available for conducting program experimentation. However, significant institutional barriers prevent their deployment. If the very significant uncertainties about program effects that surround net-to-gross energy savings and spillover effects are to be resolved, program administrators and regulators must support experimentation and the evaluation of such studies. To do so, they will need to develop the funding, manpower, and management capability to provide the proper environment for rigorous experimentation to maximize program success. This paper explains why this change in program design and evaluation is necessary and points the way toward removing some of the significant barriers that are present.

Experimental Design

Threats to Validity

Social experiments related to energy efficiency programs should be designed to more conclusively determine whether policy changes or energy efficiency program design features cause the desired changes in energy consumption. To do so, it is necessary to minimize the impacts of threats to internal and external validity in experiments. This is the primary objective of experimental design (see Campbell 1969 and 1988; Cook and Campbell 1979; Cook and Shadish 1994; Shadish, Cook and Campbell 2002).

Internal Validity. Internal validity describes the validity of inferences (or conclusions) that are drawn about the relationship between cause and effect observed in an experiment. Threats to internal validity are aspects of the design of an experiment that can cause experimenters to draw erroneous inferences or conclusions from the outcome of the experiment. There are many threats to internal validity. Experiments generally involve the comparison of what happens when a treatment (e.g., a policy change or energy efficiency program design feature) is present with what happens when it is not. Observing what happens when the treatment is not present is harder than it sounds on the surface. In essence, one must observe what would have happened if the experimental factor was not present – the so-called counterfactual condition. One simple way to do this is to observe subjects before and after exposure to the treatment. This approach turns out to be fraught with peril. When we make such a comparison, it is possible for a variety of alternative explanations to actually account for any difference that we observe before and after treatment. Important alternative explanations include: *History* (some factor unrelated to the treatment may have caused the apparent change in the dependent variable of interest), *Maturation* (the normal aging process may be responsible for the observed effect), *Testing* (the measurement process itself may cause the dependent variable), *Instrumentation* (calibration of the measuring instrument can slip) and *Regression to the Mean* (sample-to-sample variation may look like change).

An alternative and complementary approach to before-after measurements is to measure the effect of a treatment variable by observing the difference between statistically identical groups of subjects – one that has been exposed to the treatment and another that has not. Of course, the validity of such a comparison rests on the assumption that these so-called treatment and control groups were identical in all meaningful respects before exposure to the treatment factor – so that any observed difference is the result of the treatment. If they are not identical, then you have what is known as a selection effect – probably the most pervasive threat to the internal validity of social experiments.

External Validity. The external validity of an experiment refers to whether or not the results obtained in a given experiment can be generalized from the circumstances of the experiment (the study groups) to a broader set of circumstances (e.g., the population of residential customer households). There are three major threats to external validity. If the *subjects* observed in an experiment (e.g., students in a dorm) are significantly different from those of the population for which the generalization is to be made (e.g., residential households), there is reason to suspect that the causal relationship observed in the experiment may not occur for the population of interest. It is also possible that the *setting* (e.g., commercial buildings) to which the generalization is to be made is very different from the setting in which the experiment was conducted (e.g., hotels), and that the causal relationship observed in the experimental setting will not be true of the situation to which the experimental result is expected to be generalized. Finally, if the *treatment or outcome measures* are changed significantly, there is reason to doubt whether the causal relationship observed during the experiment will hold.

Establishing experimental procedures that ensure both internal and external validity is a critical requirement in experimentation. Experiments that are not internally valid (i.e., methodologically flawed) are generally not useful, because they do not conclusively show that the experimental variable actually causes the change in the outcome variable (e.g., kWh usage) of interest. Such experiments are, at the minimum, a waste of time and money. They can lead to more damaging outcomes if the results confirm some prior expectation, and therefore, are readily accepted without additional verification.

In designing experiments involving humans, it is important to keep in mind the fact that there are often tradeoffs between the risks imposed by internal and external validity. For example, in creating a robust RCT, the experimental setting may become so dissimilar to the real world that the usefulness of the experiment for extrapolating to real world conditions is undermined (as noted above in the discussion of external validity and students in dorms).

Concern about controlling the threats to internal and external validity has led to calls for greater use of RCT designs in assessing the impacts of energy efficiency program design alternatives (e.g., Allcott and Mullainathan 2010). RCTs compare the outcomes for groups that have been randomly assigned either to treatment groups or to control groups before the intervention. This approach to constructing comparison (i.e., treatment and control) groups logically and mathematically eliminates most of the threats to internal validity that can account for observed differences between treatment and control groups on outcome measures – provided sample sizes are reasonably large and selection does not occur *after* the assignment to experimental conditions. An observed difference in outcome measures for groups formed in this way will generally provide a robust measurement of program impact.

While random assignment is the most robust approach to assessing program effectiveness, it is not the only rigorous research design available and is not always feasible, as noted below (Sullivan 2009). Quasi-experimental designs are often the only recourse available to researchers operating in applied research settings. Nevertheless, the movement away from RCT designs to quasi-experiments is a slippery slope, and researchers must recognize that the compromises that are made when moving to these designs may render their conclusions indefensible. The history of social experimentation is replete with examples in which seemingly promising findings obtained in quasi-experimental studies in medicine, education, welfare and employment have been overturned in subsequent, more definitive experiments involving RCT designs (Coalition for Evidence-Based Policy 2009).

Experimentation (particularly randomized experiments) has been conducted for many years in a number of fields outside of the energy efficiency arena: e.g., public health (anti-drug, anti-smoking, pharmaceuticals, medical devices), education, social services, media, military, etc. (Duflo et al. 2007; Megdal and Bender 2006; Greenberg and Shroder 2004). Greenberg and Shroder (2004) document 241 completed and 21 ongoing social experiments, all of which include random assignment. Movement toward *evidence-based* programs, policies and interventions is possible because of the much richer research traditions – and particularly experimental research – in those fields.

Barriers to Conducting Experimental Designs

Many researchers responsible for evaluating energy efficiency programs are trained in research methods and are generally aware of the benefits of RCT designs. Yet, these research designs are not widely used in assessing the impacts of energy efficiency programs and policy changes. This is because there are several significant barriers to the widespread use of RCT designs and experimentation in general including: regulatory, institutional, design and scope/theory.

Regulatory Barriers. Regulatory barriers often prevent the use of experimental design. First, intervenors often resist some of the key requirements for experiments. For example, they sometimes argue that anything offered to any customer has to be (simultaneously) available to all customers; that some customers must be protected from interventions that might harm them economically; and that nobody can be worse off as a result of exposure to an experimental treatment. All of these arguments can lead to the imposition of constraints that make meaningful experiments of programs or policies under consideration impossible. Second, regulators are often impatient and desire inexpensive studies that can be done quickly (i.e., they don't want to wait months or years until the results from experimental studies are available). Consequently, evaluation designs often favor approaches to impact assessments based on engineering estimates or stipulated savings. Finally, regulatory staff are sometimes not familiar with the technical advantages arising from the use of experiments and the requirements associated with carrying them out correctly and do not press utility evaluators to use such designs to address critical issues.

Institutional Barriers. Institutional barriers also seriously constrain the use of experimental design in assessing the effectiveness of program design elements and program impacts. First, program administrators generally do not have corporate experience with experimentation (particularly with their customers), and their evaluators will meet some resistance to doing experiments simply based on the fact that experimentation hasn't been widely used to improve program performance in the past. Second, most program administrators are not experienced in designing and conducting experiments, nor in the management of the benefits and costs of experimentation. This is a particularly serious problem with large organizations where various departments (e.g., marketing, billing, strategic planning, generation planning, distribution planning and senior management) are all involved in program planning and implementation. Those responsible for measurement and evaluation may advocate the use of well developed experimental designs for assessing program effectiveness, but they generally do not have control over many aspects of implementation (e.g., marketing or customer services) and, therefore, must compromise with other departments whose interests may not coincide with the development of good scientific investigations. In many cases, it takes only one person to completely undermine the effort. Third, utilities are generally unwilling to force customers into treatments or withhold treatment randomly, fearing customer backlash and for equity concerns.[2] It is almost never the case that customers are put in a treatment from which they cannot escape, and as a result, there is usually an issue of selection bias in comparing treatment and control groups regardless of whether they were initially randomly assigned. Fourth, program administrators are risk averse and by inertia alone are resistant to putting long-established relationships and understandings at risk. Finally, because experimental results can be dispositive of the issue of the effectiveness of programs, program administrators risk the loss of incentive payments and even cost recovery when robust experimental designs are used to assess program impacts. The bottom line is that experiments pose a lot of risks to parties involved in energy efficiency program development and operation, and the benefits cannot easily be demonstrated before the experiments are carried out.

Design Barriers. Design barriers may limit the use of appropriate experimental designs. In addition to the above concerns shared by regulators and program administrators, statistical experiments like RCTs require large enough samples to rule out the possibility that the observed differences between treatment and control groups could have occurred by chance alone. With small samples, there may be such large sample-to-sample variation in outcome measures so as to make it impossible to detect a statistically meaningful effect. Moreover, with small samples, it is also possible that treatment and control groups may not be statistically "identical" as a result of random assignment. In other words, other antecedent causes of the variables may not be randomly distributed across the comparison groups, so that we won't really know whether an outcome is really a treatment effect or caused by something else. With repeated trials and large enough samples, it is possible to rule out the possibility that a program's 2% effect was really non-existent or not statistically distinguishable from, for example, a 1% difference. With small effects (such as a 2% response typically observed with normative feedback programs where a household's energy use is compared to that of similar neighbors) (Alcott 2011)), relatively large samples are required in treatment and control groups to have any confidence that the 2% is not a statistical artifact (random outcome), and even then it is prudent to wait for such an effect to be replicated before

[2] Universities have established Federally mandated human subject principles and procedures that assure informed consent and good risk/benefit ratios for academic research. These could be a stumbling block to the design of experiments if universities were involved.

pronouncing the approach a success. If the agency or regulator is not ready to commit sufficient resources for large samples, then the randomized experiment may not be viable.

Scope/Theory Barriers. Scope/theory barriers may make the use of RCT designs impractical or impossible. For this paper, scope refers to the underlying theory of how the program is hypothesized to impact the population. RCT designs are particularly suited to measuring impacts in situations in which the program initiative is supposed to *directly* affect the behavior of the subjects. In such cases, it is reasonable to infer program impacts by observing changes for individuals that have been subjected to the treatment condition. This is an important class of energy efficiency program initiatives.

However, there are energy efficiency intitiatives that are not designed to achieve direct effects. For example, in several states, some energy efficiency programs are intended to catalyze widespread changes in behavior and in the market by changing the availability and price of energy efficient products. Such "market transformation" programs are typically aimed at the market – not the individuals who comprise it. It is really only possible to observe the effects of such programs by comparing markets that have been exposed to such conditions with those that have not been exposed on critical market level indicators (e.g., availability of energy-efficient products). It is generally not possible to control the presentation of such market level interventions (i.e., to randomly assign markets to treatments). So, quasi-experimental methods are probably the only recourse for studying the effects of such initiatives.

Spillover poses similar though less serious problems. Spillover occurs when a program that is targeted at some subset of the population can cause a change in energy use for some other subset of the population that was not treated. This is what is called an *indirect* effect. Unlike market transformation initiatives, spillover can be directly studied using RCT designs under some circumstances. However, doing so will require a very careful experimental design that randomly varies populations (e.g., neighborhoods or social networks) that are exposed to the treatment and then within those social groups randomly varying exposure to the treatment. This approach has been effectively used in public health (Miguel and Kerner 2004).

Overcoming Barriers to Implementing Experimentation

We recommend the following changes in the regulatory and institutional environment to promote the use of experimental design. First, regulators and intervenors must allow experiments that result in situations where benefits may not accrue to all customers simultaneously, or even at all. Second, regulators must recognize that well designed experiments often require time. Third, regulatory staff's understanding of and familiarity with the technical requirements associated with experimental design must be improved, either by providing training courses in experimental design to existing staff or hiring staff with these capabilities. Fourth, regulators should reward program administrators for conducting useful experiments; otherwise, program administrators will under-invest in experiments as they are public goods where all program administrators benefit from the lessons learned from one experiment run by one program administrator. Fifth, regulators should consider hiring experts in experimental design to be on staff to help program administrators for implementing experiments.[3]

Solutions also must be found for eliminating institutional barriers. First, program administrators must be *encouraged* to employ experimentation as a means for innovation (Sullivan 2009). This is less a matter of changing the corporate cultures of organizations responsible for program implementation than it is a matter of demanding innovation on the part of these parties. "Reinvention laboratories" (Ehrhardt-

[3] Energy efficiency may not be sufficient to justify this, but the challenges raised by introducing new technologies, many associated with the "smart grid," provide other situations well-served with good experimental design.

Martinez and Laitner 2009) may need to be created where space and flexibility are provided to innovate and experiment with specific energy efficiency strategies and programs. Second, program administrators must fund the staffing and other resources required to provide the proper environment for rigorous experimentation designed to maximize program success. Third, program administrators, with regulatory approval, must be willing to randomly expose customers for treatment and control groups and develop the experience required to manage customer backlash. This will most likely require customer education as well. And fourth, program administrators must be given the economic incentives to accept more risk of failure to meet goals or other performance metrics. In the current regulatory environment, program failure, which can occur quite easily when experimenting with new ideas, can result in significant economic losses of either program costs, performance incentives, or both. Paradoxically, the incentive systems that are in effect in some jurisdictions are severely restricting innovation in program development because the risk of failure in experimentation is taken into account by program administrators when they decide on program design features, and maximization of incentive payments is so heavily weighted by the program administrators in the program design process. (For more discussion on the interaction between evaluation and incentives, see Blumstein 2010).

Finally, it may be prudent to develop protocols or guidelines for conducting experiments, especially to help those without any experience in experimental design. While the academic literature is replete with this type of guidance information, there is very little available in the energy efficiency arena. The guidelines developed by the Electric Power Research Institute (2010) for energy information feedback programs are a good start.

Alternative Experimental Designs

In true experimental designs, participants (e.g., households or businesses) are randomly assigned to treatment and control conditions. These designs are definitely to be preferred over the less robust alternatives discussed below, and every effort should be made to adhere to sound conventional designs based on randomization. Nevertheless, the kinds of practical considerations discussed above will often make use of true experimental designs impossible, and therefore, it is often necessary to employ practical, second-best alternatives called quasi-experimental designs.

Quasi-experimental Designs. There are several types of quasi-experimental designs that vary according to their robustness (i.e., the extent to which they can achieve the credibility of a true experiment) and difficulty in execution. They are: (1) regression discontinuity designs, (2) non-equivalent control groups designs, (3) interrupted time series designs, and (4) randomized encouragement designs (Duflo et al. 2007; EPRI 2010, Sullivan 2009). The first three designs are described in detail in EPRI (2010) and Sullivan (2009). The fourth design is not in these references and is described in the next paragraph.

The randomized encouragement design (RED) can be used in situations with little control over the assignment of subjects to the experimental conditions of interest (Duflo et al. 2007). For example, situations in which subjects either volunteer for exposure to the treatment or can easily avoid it are very common characteristics of programs designed to improve energy efficiency. The key idea behind the RED design is that instead of randomizing the application of the intervention itself, what is randomized is *encouragement* to receive the treatment. By randomly assigning subjects to different levels of encouragement and carefully tracking outcomes for all those who do and do not receive the encouragement, it is possible to obtain reliable estimates of *both* the encouragement and the intervention itself. Encouragement can encompass a wide range of offerings, such as training, additional direct mailings, monetary incentives, etc. Note that encouragement is merely that – encouragement. Some households receiving encouragement may not follow through on the possible intervention. All that is

required is that the encouragement increases the likelihood that households will follow through with what they are being encouraged to do. It is possible to use instrumental variable analysis techniques to estimate the treatment effect for the population overall (the "intention to treat effect") and the effect of the treatment for the subset of the population that was treated (the "local average treatment effect"). The only real drawback to this approach is that weak encouragement effects may lead to situations in which very large sample sizes of encouraged and not encouraged populations are required to estimate program effects.

New Technologies and Experimental Design

New technologies in capturing data in a more timely fashion may make it easier to conduct experimentation. For example, high-resolution metering, advances in sensor/control and communication (wireless, IR, the cloud, etc.) technology, rising publicly accessible data (e.g., Google Earth), and vastly increased computation power offer opportunities that were not available until recently. These technologies may help reduce the cost of conducting experiments and provide more timely results based on field data, helping to mitigate two key barriers to experimental design.

Experimental Design for Energy Efficiency Programs

There are two important classes of energy use behavior that now are starting to be the target of experimental design: (1) actions that result in the adoption of more energy-efficient technology (e.g., as a response to marketing initiatives); and (2) practices that result in changes in energy consumption (e.g., thermostat settings, lighting controls, etc.). These are two broad classes of behavior that can significantly affect the success of initiatives designed to improve energy efficiency.

As noted in Sullivan (2009), there have been very few published experiments designed to test alternative approaches to the design of energy efficiency program delivery mechanisms, such as message content, advertising, targeting, channel effects, social network effects, or other aspects that might improve the likelihood that consumers adopt the target technologies and behaviors. While notable efforts have been underway for decades to demonstrate the efficacy of new energy-efficient technologies (e.g., higher efficiency lighting), there has been almost no systematic effort to demonstrate more effective means of causing consumers to adopt new and more energy-efficient technologies – at least no effort using the techniques commonly used in product development in business and industry. In his review of experimental design studies, Tiedemann (2011) found that the majority of field experiments relied on two theoretical perspectives: rational choice and the theory of planned behavior. He noted that the additional theoretical perspectives stemming from applied psychology and social psychology were rarely reflected in experimental studies. However, in the past two years this situation has been changing.

In recent years, considerable interest has developed in energy conservation programs that rely on the presentation of normative comparisons to encourage consumers to reduce their consumption of electricity and natural gas (Faruqui et al. 2009; Fischer 2008). The efficacy of these normative comparison approaches has been demonstrated using robust statistical experiments that more or less unequivocally demonstrate that subtle but statistically significant changes in energy use behavior can be caused by providing consumers with normative information about their energy use (e.g., comparing a household's energy use to that of similar neighbors).

During the past two years, a company called OPOWER has partnered with over 40 utilities throughout the U.S. to send energy use reports to residential electricity and natural gas consumers. The reports display the household's energy consumption, compare it to similar households over time, and provide energy conservation tips. The social comparisons are based on research that shows that

descriptive social norms are better at reducing energy use than appeals to saving the environment and to social responsibility. OPOWER's programs were designed for rigorous evaluation: from a population of households in the utility's service territory, some are randomly selected to receive the report letters, while the rest remain as a control group.

Early evaluations of the OPOWER program have focused primarily on electricity savings (e.g., Allcott 2011; Ayres et al. 2009; Summit Blue Consulting 2009), and they have shown that the programs cause households to reduce energy use by about two percent, depending on the program's location, frequency and duration. In one study, the decrease in energy use was more likely due to behavior changes (e.g., turning lights off) than physical measures (e.g., weatherization) (Ayres et al. 2009). However, most evaluations of these and similar type programs have provided very limited information on implementation challenges and on specific changes to behavior.

Finally, another area in which experimental design is being used today is in assessing the impacts of dynamic pricing on the timing and magnitude of electricity consumption (Faruqui and Wood 2008; Faruqui and Sergici 2009). The experimental design includes control groups that stay on the standard tariff and treatment groups that are placed on new time varying tariffs or information programs. The treatment groups for each tariff are often divided into subgroups that face different price levels, so that statistical relationships between energy use by rate period and prices can be estimated. Recently, the U.S. Department of Energy (DOE) began a program that will invest $3.4 billion in Smart Grid technologies to modernize the nation's electric grid, and DOE has funded a number of projects through Smart Grid Investment Grants, some of which are using randomized control experiments to study the impacts of dynamic pricing. These projects are expected to start in 2011 or 2012, as they are in the process of obtaining regulatory approvals for their study plans and rate treatments (Cappers 2011).

Recommendations for Implementing Experimentation

As noted above, experimentation can be complicated, expensive and time consuming. Thus, experiments should not be undertaken without considering the benefits and costs in terms of time and resources. The key to success in innovation lies not in applying experimentation to test all possible improvements to programs but in strategically applying experimentation to obtain answers to critical questions.

From the point of view of achieving immediate results, experiments should probably focus on those technologies with which there is the most experience (e.g., CFLs); where the demand savings may be the largest (e.g., summer air conditioning in hot places); where the total energy savings may be the largest (e.g., plug loads and standby losses in residential households); or where findings can easily be generalized to other products. The experiments could also be on packages of energy efficiency measures, or packages of energy efficiency measures and energy use behavior changes.

Experiments should be focused on those aspects of program delivery that can be manipulated to create the most leverage in changing energy consumption. These include initiatives related to the impacts of (1) information, education, knowledge and experience (by varying form, content, delivery system/messenger, frequency and duration); (2) inducements (subsidies, rebates, price breaks, ex-post rewards and praise/approval), costs, penalties and prices (by varying amount, timing, framing, recipients and delivery system); and (3) market interventions such as (a) point-of-sale delivery systems (by varying signage, advertising, packaging with other items/services), (b) mid-stream programs and (c) upstream programs (by varying education, inducements, service provisions and competition).

At the moment, the research community is heavily focused on discovering how variations in the presentation of descriptive norms and financial and environmental cost information influence energy use

behavior. While providing this kind of feedback is potentially a very productive area, it is unimaginative. Consider the fact that without being asked about the details of their lifestyle, appliances and other energy use-related factors, households are receiving comparisons that are supposed to represent the difference between their energy use and the energy use of a statistically comparable "average" household. This is a simplistic framework with lots of room for improvement. The world is evolving, so that we can improve both the design and evaluation of these programs. For example, customers could be allowed to customize comparisons, so that the comparison group is more like themselves, and software could be provided that enables customers to simulate different energy use strategies and see their effects. For example, Paul Hines at the University of Vermont is in the preliminary stages of an experiment using a randomized encouragement design to evaluate an energy efficiency web-based social network, and he and his colleagues are hoping to make their tool available to a wider audience (Hines 2011).

A very productive area that is not being addressed systematically is the use of experiments to improve the quality of marketing efforts. Experimenting with alternative marketing techniques, such as different messages, channels and strategies, is relatively easy to do, as compared to experiments designed to change "use related" behavior, and can lead to very profound advances in the effectiveness of marketing. Experimental designs for evaluating message impacts are well developed and inexpensive to carry out. Market segmentation schemes will also play a key role in understanding the effectiveness of the messages for key groups.

Another promising area is the evaluation of community level interventions. For example, local governments, schools and community groups are being used as part of social marketing initiatives to promote energy efficiency. A framework for systematic experimentation needs to be developed and used for studying effectiveness and experimenting with new ideas at the community level.

Ultimately, the range of possible experiments is limited only by social science theory and our imaginations. Thinking about more "basic" research that should be undertaken using experimentation, there is a rich storehouse of social science theory that may support the development of new program initiatives. For example, Paul Stern's multi-level choice-in-context approach (Stern 2008) suggests a focus on decision-making and information processing, but with attention to social influences and a variety of contextual factors. Experiments could attempt to affect certain factors, while controlling for others. The model would predict that different sets of factors are likely to be important for different technologies, behaviors, or outcomes. As Lutzenhiser (2009) noted in his review of the literature, several types of variables could be considered for applying to interventions. Variables could be drawn from different theories and might include considerations such as: costs, subjective assessments of outcomes, influential social norms, personal norms, level of effort required, knowledge and skill, cultural pressures, constraints, and influences in supply chains, etc. The list could also include whatever barriers to change might be identifiable. Lutzenhiser noted that not all of the variables that may turn out to be important are highlighted by social science theories, and some remain to be discovered and/or pointed out by experienced program staff.

In every case, the experiment must be carefully controlled in terms of treatment delivery, the control of other influences and confounding factors, and the careful measurement of all factors. The less control, the greater the need for measurement of other factors and the larger the needed sample in order to detect and estimate treatment effects with any degree of confidence. Careful thought is needed in the field of experimentation: while program experiments can be imagined and might be quite valuable, they are much easier to imagine than they are to plan, design, execute, analyze, and have confidence in the findings – and result in findings that are useful, powerful, and/or generalizable (Lutzenhiser 2009). A systematic way of designing experiments is needed.

Acknowledgements

We would like to thank the following reviewers of an earlier version of this paper: Hunt Allcott, Peter Cappers, Don Dohrmann, Ahmad Faruqui, Matthew Kahn, Phil Moffitt, Monica Nevius, Wesley Schultz, and Catherine Wolfram.

References

Allcott, H. 2011. "Social Norms and Energy Conservation," *Journal of Public Economics*, forthcoming.

Allcott, H. and S. Mullainathan. 2010. "Behavioral Science and Energy Policy," Boston, MA: Massachusetts Institute of Technology. This is a longer supporting version of an article in the March 5, 2010 issue of *Science* magazine.

Ayres, I., S. Raseman, and A. Shih. 2009. *Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage*. NBER Working Paper No. 15386, Washington, DC: National Bureau of Economic Research.

Blumstein, C. 2010. "Program evaluation and incentives for administrators of energy-efficiency programs: Can evaluation solve the principal/agent problem?" *Energy Policy* 38: 6232–6239.

Campbell, D. 1969. "Reforms as Experiments," *American Psychologist* 24: 409-429.

Campbell, D. 1988. "The Experimenting Society," in D. Campbell and S. Overman (Eds.), *Methodology and Epistemology for Social Science: Selected Papers* (pp. 290-314). Chicago, IL: University of Chicago Press.

Cappers, P. 2011. Personal communication with Peter Cappers, Lawrence Berkeley National Laboratory, February 25.

Coalition for Evidence-Based Policy. 2009. "Appendix II: Comments from the Coalition for Evidence-Based Policy," in U.S. General Accounting Office, *Program Evaluation: A Variety of Rigorous Methods Can Help Identify Effective Interventions*, Washington, DC: U.S. General Accounting Office.

Cook, T. and D. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago, IL: Rand McNally.

Cook, T. and W. Shadish. 1994. "Social Experiments: Some Developments Over the Past Fifteen Years," *Annual Review of Psychology* 45: 545-580.

Duflo, E., R. Glennerster, and M. Kremer. 2007. Using Randomization in Development Economics Research: A Toolkit. London, UK: Centre for Economic Policy Research.

Ehrhardt-Martinez, K. and J. Laitner. 2009. *Pursuing Energy-Efficient Behavior in a Regulatory Environment: Motivating Policymakers, Program Administrators, and Program Implementers*, Berkeley, CA: California Institute for Energy and Environment.

Electric Power Research Institute [EPRI]. 2010. *Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols*, Palo Alto, CA: Electric Power Research Institute.

Faruqui, A. and L. Wood. 2008. *Quantifying the Benefits of Dynamic Pricing in the Mass Market*. Washington, DC: Edison Electric Institute.

Faruqui, A. and S. Sergici. 2009. *Household Response to Dynamic Pricing of Electricity: A Survey of the Experimental Evidence*. Washington, DC: The Brattle Group, Edison Electric Institute and the Electric Power Research Institute.

Faruqui, A., S. Sergici, and A. Sharif. 2009. *The Impact of Informational Feedback on Energy Consumption: A Survey of the Experimental Evidence*. Washington, DC: The Brattle Group.

Fischer, C. 2008. "Feedback on Household Electricity Consumption: a Tool for Saving Energy?" *Energy Efficiency* 1, 79–104.

Greenberg, D. and M. Shroder. 2004. *The Digest of Social Experiments (3rd ed.),* Washington, DC: Urban Institute Press.

Hines, P. 2011. Personal communication with Paul Hines, University of Vermont, May 1.

Lutzenhiser, L. 2009. *Behavioral Assumptions Underlying California Residential Sector Energy Efficiency Programs*. Berkeley, CA: California Institute for Energy and Environment.

Megdal, L. and S. Bender. 2006. "Evaluating Media Campaign Effectiveness: Others Do it Why Don't We?" *Proceedings of the 2006 Summer Study on Energy Efficiency in Buildings*. Washington, DC: American Council for an Energy-Efficient Economy.

Miguel E. and M. Kerner. 2004. "Worms: Identifying Impacts on Education and Health on Treatment Externalities," *Econometrica*, 72(1): 159-217.

Shadish, W, T. Cook and D. Campbell. 2002. *Experimental and Quasi-Experimental Design for Generalized Causal Inference*, Houghton Mifflin.

Sullivan, M. 2009. *Using Experiments to Foster Innovation and Improve the Effectiveness of Energy Efficiency Programs*, Berkeley, CA: California Institute for Energy and Environment.

Summit Blue Consulting. 2009. *Impact Evaluation of OPOWER SMUD Pilot Study*. Boulder, CO: Summit Blue Consulting [now called Navigant Consulting].

Tiedemann, T. 2011. "Behavioral Change Strategies That Work: A Review and Analysis of Field Experiments Targeting Residential Energy Use Behavior," Chapter 21 in K. Erhardt-Martinez and S. Laitner (Eds), *People-Centered Initiatives for Increasing Energy Savings*, E-Book, Washington, DC: American Council for an Energy-Efficient Economy.