Using Experiments to Foster Innovation and Improve the Effectiveness of Energy Efficiency Programs

Prepared by
Michael J. Sullivan, Ph.D.
Freeman, Sullivan & Co.

Prepared for:
CIEE Behavior and Energy Program
Edward Vine, Program Manager

Prepared for the
California Institute for Energy and Environment and the
California Public Utilities Commission's Energy Division

March 2009

DISCLAIMER

This report was prepared as an account of work sponsored by the California Public Utilities Commission. It does not necessarily represent the views of the Commission or any of its employees except to the extent, if any, that it has formally been approved by the Commission at a public meeting. For information regarding any such action, communicate directly with the Commission at 505 Van Ness Avenue, San Francisco, California 94102. Neither the Commission nor the State of California, nor any officer, employee, or any of its subcontractors or Subcontractors makes any warranty, express or implied, or assumes any legal liability whatsoever for the contents of this document.

Table of Contents

Figures

Executive Summary

Increasing energy efficiency is a cornerstone in the effort to reduce greenhouse gas (GHG) emissions and lower America's dependence on oil supplied by parties under the control of foreign governments. It has been estimated that no more than 60% of the economically justifiable energy efficiency potential is being achieved in California at this time. With time running out on the opportunity to avoid the potentially catastrophic societal and economic consequences of our current level of reliance on hydrocarbon fuels, the challenge we face is to capture that remaining 40% and more of energy efficiency potential as soon as possible.

Much of the remaining economically justifiable energy efficiency potential lies in changing the *behavior* of energy users – in particular, changing the decisions that they make about adopting energy-efficient technologies and practices, and in changing the ways in which they use energy (e.g., lifestyle changes). A review of the literature in psychology, sociology, social psychology and behavioral economics suggests that some behavioral science-based approaches to improving the acceptance of energy-efficient products in the market show great promise. However, this promise has yet to be realized because practical field experiments are required to discover what works and what does not work and why, and these experiments have not been conducted.

To improve the effectiveness of energy efficiency programs by increasing the likelihood that consumers adopt energy efficient technologies and practices, a formal research and development (R&D) effort designed to find effective strategies for improving energy efficiency program performance must be undertaken. This effort should focus on discovering effective behavioral science-based strategies for improving the performance of existing programs and on developing new and more effective approaches to offering these programs. Currently, California government and regulators sponsor substantial R&D designed to accelerate the rate at which more energy-efficient technology is available in the market. At the same time, almost no R&D is expended that is intended to improve the likelihood that customers adopt these technologies once they are commercially available. This is a significant gap in program development.

There are well-established procedures for managing and carrying out product and service R&D efforts. They are generally discussed in the academic literature under the heading of the Management of Innovation. Innovation is managed by moving new product and service design ideas through a stepwise process from idea generation at the very beginning, to full-scale integration with business operations at the end. Along the way, a number of appropriately scaled experiments are carried out to solve the myriad technical problems that surround the development of something new. Most major corporations rely on innovation to survive and to maintain their market position. This paper argues that the establishment of a process designed to manage innovation must be developed in California to foster the creation of badly needed program improvements and develop new and more effective energy efficiency delivery programs.

Experimentation is a critical requirement in the process of innovation. It is the mechanism that innovators use to identify what works and what does not work during the process of product development and marketing. Historically, there is very little evidence of the use of experimentation to test alternative energy efficiency program design features offered by utilities in California or elsewhere. Instead, programs tend to emerge full-blown from concept testing to implementation – without significant prototype development and testing.

Pilot studies and demonstration projects have limited usefulness in product and program development – particularly when they are undertaken before the program has undergone prototype testing. Because pilots are generally large-scale program demonstrations and do not vary design alternatives, they don't produce useful information for improving program performance. It takes months or even years to complete a pilot in the current energy efficiency program environment and, in the end, very little can be learned. This paper argues that instead of pilot testing, realistic small-scale experimental versions of key program components (i.e., messages, delivery channels, social network effects, etc.) should be completed prior to any full-scale pilot testing.

In part, the rush to market of energy efficiency programs is driven by the perceived urgent need to implement energy efficiency programs. Probably more important are significant institutional barriers (inside utilities and in the regulatory relationship) that discourage experimentation. This paper argues that the need to develop effective programs outweighs the need to act quickly and that institutional barriers are surmountable.

To stimulate interest and thought about how experimentation can be used to improve program performance, this paper describes a number of experimental techniques that can be applied to the study of the impacts of behavioral factors on consumer decision-making. It provides examples of important research questions that can be answered using experimental techniques. It is designed to provide a good working introduction to the use of experimental techniques in program development. It begins with an introduction to the Classical experimental designs (i.e., the Completely Randomized Design, the Randomized Blocks Design, the Factorial Design, and the Covariance Design). These designs are useful for illustrating the value of experimentation and the logic that underlies it. However, sometimes, these designs will be inappropriate or impossible to achieve in the context of the policy debate or organization in which the experiment is to be carried out. Therefore, in addition to the Classical experimental designs, the paper discusses several useful Quasi-Experimental designs (i.e., the Non-Equivalent Control Group Design, the Interrupted Time Series Design, and the Regression Discontinuity Design).

There is a pressing need to develop more effective methods for increasing the likelihood that consumers will adopt more energy-efficient technologies and behaviors. There are reasonable hypotheses from the behavioral sciences about how improvements can be achieved. The methodological tools are available for moving these hypotheses from the theory stage to the stage of practical application through formal organizational processes

and scientific methods that will ensure efficient use of resources and probable success. What is holding us back?

This paper discusses several key institutional problems that must be overcome to achieve significant progress, such as:

1. The need for improvement in the effectiveness of energy efficiency programs through the development of more effective behavioral interventions has to be recognized by the policy community (i.e., the California Public Utilities Commission (CPUC), the California Energy Commission (CEC), and the program design community in the utility and consulting sectors).

2. A thoughtful decision has to be made by regulators concerning the proper locus of responsibility for overseeing the necessary R&D efforts (i.e., the utilities, the University of California, a state government agency, a national energy research lab, or some combination of the above) aimed at improving the performance of energy efficiency programs through effective behavioral interventions.

3. Funding for R&D designed to improve the design of energy efficiency programs must be made available by regulators to the party(s) that oversee development of improved energy efficiency programs with the understanding that these program development costs are a necessary cost of managing and operating these programs.

4. Failures in the R&D process must be recognized as a natural part of progress in R&D efforts.

5. To the extent that utilities are assigned responsibility for such R&D, they will need to develop the manpower, management capability and business strategies that incorporate routine R&D related to improving program effectiveness into their energy efficiency program operations.

6. Regulators will have to develop the manpower, management capability and business strategies that will allow them to provide meaningful oversight of the R&D process without hindering progress.

The above institutional problems should not be taken lightly. To solve them, it will be necessary to completely revisit the way that energy efficiency programs are developed and supported. It will require major modifications to the practices that are used to manage these programs both inside and outside utilities.

Any significant R&D effort will require significant economic investment, above and beyond the cost of delivering energy efficiency programs today. At the moment, an act of faith is required to accept the notion that significant expenditures on program development based on behavioral science theories will dramatically improve energy efficiency. The effectiveness of behavioral strategies for improving the likelihood that consumers adopt energy-efficient technologies and practices simply remains to be demonstrated. This does not mean there aren't excellent reasons for believing that these

strategies will dramatically change the landscape of energy efficiency programs. Consider this fact: In 1965, the prevalence of cigarette smoking in the US was about 51% of the adult population. In 2007, the prevalence of cigarette smoking in the US was about 23% of the adult population. This is a 50% reduction in prevalence (of a behavior that is probably much more difficult to attenuate than decisions about energy-related behavior) over a 42 year period – or a reduction of about 1.2% per year. Behavioral interventions figured prominently in the achievement of this reduction. It seems likely that such interventions will prove to be even more effective in changing energy use behavior.

1. Introduction

Increasing the efficiency of energy use is a cornerstone in the effort to reduce greenhouse gas (GHG) emissions and lower America's dependence on oil supplied by parties under the control of foreign governments. While energy efficiency programs have been evolving since 1973 in the US, and a lot of progress has been made, it has been estimated that less than 60% of the economically justifiable energy efficiency potential is being achieved (Rufo and Coito 2002). Moreover, potential studies typically focus entirely on the potential from improving the efficiency of energy use in buildings and production processes. These studies completely ignore energy use arising from decisions made by businesses and households about the contents of their supply-chains and energy efficiency improvements that can come from changing lifestyles. In other words, the untapped potential of energy efficiency is probably very large. With time running out on the opportunity to avoid the likely catastrophic societal and economic consequences of our current reliance on hydrocarbon fuels, the challenge that we face is to capture that potential as quickly as possible.

Much of the remaining economically justifiable energy efficiency potential lies in changing the *behavior* of energy users (Lutzenhiser 1993; Lutzenhiser et. al., 2002, 2003; (Golov and Eto 1996; Nadel and Geller 1996; Sanstad and Howarth 1994; and Sullivan 2009). Unfortunately, while achieving significant improvements in energy efficiency may depend to a very large extent on our ability to influence important decisions made by consumers, our knowledge of how to impact consumer behaviors related to energy efficiency remains very limited.

Our inability to impact important consumer behaviors stems not from a lack of interesting theories about how to alter consumer behavior, but from a lack of practical experience in applying these theories to changing consumer choice behavior related to energy efficiency. The behavioral science literature contains a number of powerful theoretical perspectives that suggest how changes in consumer behavior related to energy efficiency can be achieved. Unfortunately, because there has been little experimentation with the application of these theories to modifying consumer choice behavior related to energy efficiency, we don't know what will work and what will not work.

To name just a few interesting observations about behavior related to energy efficiency decisions, behavioral scientists tell us:

1. Decisions are shaped by many considerations beyond perceived costs and material benefits, and appeals to motivations other than self-interest can be successful in changing behavior (Biggart and Lutzenhiser 2007). Therefore, it should be possible to develop alternative messages to consumers designed to appeal to core values, opinions and attitudes that may be much more likely to trigger desired choice behavior than information concerning costs and benefits.

2. Decision-making heuristics other than economic rationality often guide consumer and business decision-making (Sullivan 2009). Examples of such alternative

decision-making heuristics include altruism, bounded rationality, conformity and elimination by aspects.[1]

3. Feedback to consumers on the consequences of their actions can enhance the likelihood that they engage in energy-efficient behavior (Wenett et al. 1978).

4. Acceptance of new technology is strongly influenced by social processes involving the development of "shared" knowledge, opinions, beliefs and social norms on the part of consumers (Rogers and Shoemaker 1971). Correspondingly, changing consumers' perception of what is normative can strongly influence behavior related to energy efficiency.

5. People weigh the value of information received from their peers more heavily than they do others in evaluating new technology (Darley and Beniger 1981) – thus, information diffusing through a social network is much more likely to be perceived and acted on than broad spectrum media campaigns. Therefore, social marketing techniques involving infusing information into social networks through social opinion leaders and others should impact behavior related to energy use.

6. People act to reduce cognitive dissonance when they receive information that is inconsistent with their current behavior either by rejecting the information or changing their behavior (Festinger 1957; Kantola et al. 1984). Therefore, messages can be designed that are more likely to induce behavior change related to energy efficiency decision-making once customers' core values or strongly held beliefs are understood.

7. The opinions and beliefs of the management of organizations strongly determine the receptiveness of the organization to the adoption of energy-efficient technologies and practices (Lutzenhiser et al. 2003). Therefore, it should be possible to improve the likelihood that organizations make energy-efficient decisions by implementing programs designed to influence the actions of the leadership.

8. The opinions and beliefs of the management of organizations are formed and maintained in social networks that have connections inside and outside their organizations (Chattopadhyay et al. 1999).

---

[1] Consumers often apply different decision-making rules that do not appear to correspond well with maximizing benefits or minimizing risks. For example, consumers sometimes choose alternatives that improve the condition of others at their own expense (altruism) or alternatives that express their similarity to others in society (conformity). For a detailed discussion of these decision making heuristics, see Sullivan (2009).

The above partial list of interesting insights from behavioral and economic sciences suggests that lots of interesting possibilities for altering consumer behavior related to energy use, such as:

1. Discovering effective advertising and marketing strategies that cause consumers to invoke other frames of reference (other than cost benefit analysis) in making decisions involving the purchase of energy-efficient technology alternatives and lifestyles.

2. Identifying program targeting strategies that enhance the likelihood that contacted parties in the mass market act on the information they receive.

3. Developing effective community and interest group level interventions that use mechanisms in the social environment (i.e., social networks, laws, norms, and social influence hierarchies) to cause consumers to make more efficient choices.

4. Finding more effective means of fostering energy efficiency improvements by educating customers through alternative information channels (e.g., social networks, community groups, schools, etc.).

5. Measuring and documenting the effects of programs designed to foster energy efficiency by providing feedback to consumers regarding the costs, benefits and consequences of their behavior (i.e., using modern technology to assist consumers in identifying appropriate behavioral changes).

6. Measuring and documenting the effects of programs designed to achieve energy efficiency improvements by causing persistent changes in attitudes, opinions, norms and values of consumers.

7. Developing effective means for causing the management of enterprises to establish formal energy efficiency improvement goals, appoint a trusted team member to achieve them, and provide funding to support the efforts of the people in the organization to identify and implement energy efficiency improvements.

All of the above ideas make sense and are supported by reasonable empirical evidence that suggests that they may be very useful in increasing the likelihood that consumers will adopt energy efficiency technologies and practices. Unfortunately, at the present, most of these ideas are, at best, reasonable hypotheses. The problem is that little systematic effort has been made to move these ideas from the theoretical stage to the operational stage. They are good theories, but today that is really all they are. It may well be that actions implied by these theories can strongly impact the adoption of energy-efficient technologies and behavior, but we simply aren't sure how to do it.

The objective of this paper is to describe R&D strategies and experimental designs that are appropriate for discovering ways of changing specific, well-defined behaviors related to the use of energy in buildings (both residential and commercial) and production processes. In particular, the paper suggests R&D strategies and procedures for developing or modifying energy efficiency programs offered by utilities, so that they

increase the likelihood that consumers will purchase energy-efficient devices and change their energy use behavior.[2] This paper argues that innovation is needed in the development and modification of energy efficiency programs focused very explicitly on increasing the likelihood that consumers adopt the technologies that these programs offer. The essence of innovation in new product and service development is experimentation. It is at the core of all successful efforts to develop and present new products and services to the market (Thompke 2003).

A fundamental barrier to innovation in the development of energy efficiency programs offered by utilities is that there isn't an institutional framework within which program improvements can be operationally tested and implemented. Given the foci of the government agencies (e.g., the California Energy Commission (CEC) and the California Public Utilities Commission (CPUC)) and other institutional players (e.g., utilities and the U.S. Department of Energy (DOE)), there currently aren't any institutional mechanisms for systematically improving our knowledge of how to alter consumer behavior related to the adoption of energy-efficient technologies in meaningful ways.

Proposing that utilities suddenly start using innovation management techniques as described in this paper to improve the design of energy efficiency programs will not change the situation. It won't change the situation because the regulatory environment in which the utilities are operating is not currently designed to support this activity. To change the situation, the existing regulatory environment has to be modified. In closing, this paper suggests changes to the regulatory environment of utilities that would encourage, indeed require, innovation in the development and implementation of energy efficiency programs.

Before we undertake this discussion, we briefly describe how experimentation is typically used by industry to develop and market new products to consumers. Along the way, we discuss different types of market experiments and the applicability of these techniques to the development of energy efficiency programs.

2. Managing Innovation

Innovation happens when new and revolutionary ideas are translated into new products and services. The product is only part of the innovation. Innovation requires the penetration of the product into the market. It is a revolutionary process whereby an old way of doing things is replaced by a new one. Innovation doesn't usually happen by

[2] To be sure, there are other interesting consumer choices and behaviors that may have an even greater impact on societal energy use for which means could be found to alter them. However, the sponsors of this paper have an abiding interest in changing consumer behavior related to the choice of energy using equipment in buildings and business process, so the work in this paper is focused on that aspect of the problem. It is probably worth considering how other behaviors might be altered and how institutional mechanisms could be altered to foster R&D that would accomplish such research, but that work is for another day.

chance. It happens through a painstaking process of experimentation in which a better way of doing things is found by trial and error.

Perhaps the greatest innovator of all time was Thomas A. Edison. His laboratories developed the first practical electric light bulb, the 110 volt AC motor, phonograph and movie projector. In all, he was responsible for more than 1,000 patents. Prospero, as he has been called, measured the productivity of his laboratory not by the number of inventions it made or by their economic worth.[3] He measured the productivity of his laboratory by the number of experiments it completed in a day.

Experimentation is at the very core of innovation. It is how humans learn what works and, more importantly, learn what doesn't work. It is the principal means by which all progress is made. The value of experimentation is not confined to the development of things like light bulbs, computers and automobiles. It applies to the development of innovative techniques for doing all kinds of activities – activities like surgery, accounting, law, finance, marketing and sales. It also applies to the development of innovative techniques for improving the likelihood that consumers buy energy-efficient products and changing the behavior of humans and organizations related to their use of energy.

Since the early 1980s, when industry in the US began to lose its dominance in world markets for manufactured consumer goods, a substantial amount of attention has been paid in the academic literature to studying the management of innovation. There are a few seminal ideas from that literature that should be applied to thinking about the management of innovation in developing more effective energy efficiency programs. The most important idea in the literature on the management of innovation is that innovation can be managed and that it must proceed through a well-understood process that all parties involved in agree on.

[3] Prospero was the protagonist in Shakespeare's *The Tempest* who became the wizard of light. Edison was given this nickname by his friends and biographers.

2.1. Innovation Must Be Managed

The first and most important idea from the management of innovation literature is that innovation should proceed through an orderly process in stages – from idea generation to full scale implementation. The basic idea is that the R&D process should take maximum advantage of information available as early as possible to avoid making mistakes later in the process which are likely to be costly as the development process proceeds. Organizations that manage innovation typically establish formal procedures and review processes through which new product and service ideas must pass on the way to market. The design of innovation management systems varies from organization to organization, but they generally involve a stepwise process in which development moves from relatively inexpensive and quickly executable steps in product development to more expensive steps that entail much greater risk and commitment to proceed. The following is an example of an idealized product or service development process:

1. Concept development – in this stage, the proposed product or service is developed to the point that it can be described to management and potential consumers. The effort in this phase is focused on making the basic idea as concrete as possible in terms of how it actually will work, who will buy it, what they will do with it, how large the market is for it, what will be required to deliver it, and how much it will cost. At the end of this stage, the decision is made whether to take the product or service to the next stage.

2. Concept testing – in this stage, the concept is presented to customers to get their reactions to the idea and their recommendations for how it can be improved. Concept testing can be performed using a wide variety of techniques such as focus groups, in-home interviews and product clinics. The type of test determines the methods used. At this stage, the product concept may be significantly modified in response to customer reactions, or the product development effort may be stopped altogether based on what customers say about it.

3. Business case – at this stage, a business case is developed for the product or service which takes account of the information contained in the concept development and testing phases. The business case is usually presented to a review committee that has responsibility for evaluating proposals for new products or services and makes the determination as to whether the proposed business meets the criteria (risks, development costs, hurdle rates, etc.) necessary for proceeding to the next and much more expensive steps in the development process.

4. Product development – in this stage, the work required to actually make prototype versions of the product or service is carried out, and the performance of the prototype is "bench tested" to make sure it does what it is intended to do. This stage is probably the longest and most expensive activity in the development process. It is the point at which all of the technical and institutional hurdles to the development of the product must be overcome. It

is the stage when the first working version of the product is found and its efficacy is demonstrated for the first time. This stage often involves a great deal of experimentation to overcome design issues that are inherent in new products or services.

5. Market testing – once the prototype product or service has been fully developed, an effort is made to test the receptiveness of the market to the product through the distribution channels and advertising that are being considered for the launch. Techniques for market testing range widely. In the automobile industry, for example, potential new car buyers are exposed to prototype models and asked to compare them with alternatives - present models and those provided by competitors (these clinics can cost as much as $1 million per project). In packaged goods, products are often placed in a panel of stores along with facings, displays and point-of-purchase promotions like those that are being considered for use in launching the product. The performance of the product in these "mini markets" is then evaluated in relation to that of competitors and adjustments are made as necessary. The purpose of market testing is to evaluate the performance of the product under conditions as similar as possible to those in the market. The results of this evaluation are then used to adjust the product and marketing collateral for the launch. Products can make it all the way to this stage and be scrubbed, but it is not common.

6. Production – at this stage, the process needed to support production and delivery of the product is ramped up, and the product is launched.

This orderly process or something like it takes place continuously in nearly all companies that are involved in developing new products and services. Very large firms like Toyota, General Motors, Boeing, Genentech, Microsoft, Proctor and Gamble, Eli Lilly and Bank of America have numerous new products under development at any point in time, all at different stages in the development process. For many companies, this effort is vital to their survival as competition is steep and products have a limited lifespan in the context of the actions of competitors. Products are normally brought to market as quickly as possible, but not before they are ready. Nothing fails quite as spectacularly as a product that is brought to market before it is ready.

Even small companies that are starting up products in competitive industries generally use a version of this process, because the discipline of the market forces them to do so. Venture capitalists inject capital into firms that can show that they are moving their product through an orderly development process of the kind described above. Firms that cannot demonstrate progress along the above described lines are sometimes not funded; and if they are funded, they are subjected to rigorous timetables designed to move them through the above-described process.

2.2. Organizing for Innovation

Organizations that must innovate to survive in competitive markets (e.g., automobiles, consumer electronics, pharmaceuticals, software, medical devices, etc.) establish R&D departments that have primary responsibility for developing products and services through the first three stages of product development (i.e., from concept development to prototype development). During the later stages of the process, the other departments that have responsibility for production and marketing are involved.

The above process can be implemented in an organization using a variety of organizational structures. In some companies, the process is implemented within major product lines by establishing R&D departments within them. In other organizations, a central R&D department is established to serve the whole enterprise. Regardless of how it is organized, the process is implemented by establishing an R&D function that employs a systematic process to screen ideas for new products, select promising candidates based on their fit with the goals of the business, support their development, and integrate developed products back into the normal operating environment of the business.

A good example of how a major services firm used innovation management techniques to improve its business services was provided by Stephan Thomke (Thomke 2003). Prior to the 1990s, the Bank of America spent very little time and resources on R&D related to the development of its products and services. There was no formal process for developing new products and services, and, consequently, ideas for new products tended to go straight from the minds of their originators to test markets – with frequently disappointing and sometimes dangerous results. However, in the early 1990s, as a result of a variety of converging events, the company established a formal product and service development team, known as the Innovation and Development (I&D) Team. This team developed a formal process for developing and testing new products and services and implemented a "test bed" comprised of 24 bank branches located in their Atlanta market. The test bed was used to carefully experiment with a wide variety of changes to its customer service delivery systems. The objective of the test bed was to provide locations where small-scale experiments could be carried out rapidly to determine their impacts on customer behavior and satisfaction. Using this approach, the bank was able to conduct small-scale experiments on customers to discover what they liked and disliked about new products and services. This made it possible to inexpensively vet new ideas and perfect operations before large numbers of customers were exposed to them. They have developed and tested a number of branch improvements on the test bed. Because there are 24 branches in the test bed, it provides a powerful capability for conducting carefully controlled experiments on changes in all kinds of banking services on customer satisfaction and other key indicators of customers' perception of service quality.

2.3. Principles of Effective Innovation Management

Another important idea from the literature on the management of innovation is that certain basic principles should guide the development of new products and services (and, by way of inference, energy efficiency programs). These principles are as follows:

i. Someone must be assigned responsibility for knowing everything about the project – someone has to be responsible for knowing how the whole product development exercise fits together and what the status is of everything at any point in time. This person is the project champion who is responsible for representing the project to management oversight committees and boards and keeping the various parts of the product development process moving.

ii. Maximum utility should be obtained from information early on in the development process – when important problems are found late in the development of a product or service, the results can be devastating. Little problems that can be easily solved in the beginning of the product development process can become gigantic problems during later development stages. The best examples of these kinds of problems are what are called integration problems. These are situations in which two parts of a product (being designed by different parties) are occupying the same space (i.e., the parts don't fit) in the design. If these problems are resolved early on in the development of a product or service, they can be resolved inexpensively. If they are found later, they may present very serious delays, re-engineering challenges and dramatic cost overruns.

iii. Experiment frequently – experiments are expensive, but late stage failures and rework are more expensive. It is better to experiment with small-scale models of designs early and often then to wait until the end of the design and development process to perform a big test on the whole system. It is better for two reasons. First, the big test may not reveal which of the components of the system is contributing to failure; and second, small-scale tests early on may reveal problems which, if resolved at the early stage, are much less expensive to resolve.

iv. Experiment rapidly – the purpose of experimentation is to obtain relatively immediate and useful feedback. Experiments that take too long to complete may impede progress toward a solution to the problem rather than enhance it. Ideally, experiments should be completed within 90 days of commencement. There will, of course, be circumstances when the time required to complete an experimental cycle considerably exceeds 90 days. The effects of some drugs on laboratory models may take years to observe. In these situations, multiple simultaneous experiments are recommended. Grand-scale experiments requiring months or years to complete should be reserved for really significant (i.e., high risk, high benefit) long-term problems.

v. Fail early and often but avoid mistakes – failures in experimentation are to be expected and provide valuable information about the way the world works. Mistakes that happen when experiments are not properly controlled or designed provide little or no information about how the

world works and are to be avoided. At best, mistakes are wasteful. At worst, they can lead to the development of inferior products.

Taken together, the above principles define a highly controlled program development environment in which mostly small-scale experiments with program elements are performed early on in the development process in order to minimize the risk of big program failures. These design principles are well-suited as guidelines to developing R&D projects designed to discover effective behavioral science-based approaches to improving the likelihood that consumers adopt more energy-efficient technologies and practices.

2.4. Implications for Developing Energy Efficiency Programs

Most of the existing energy efficiency programs being operated by California utilities were not developed using an innovation management process of the kind outlined in the preceding sections. This conclusion is supported by several observations:

1. Most energy efficiency programs were developed by energy utilities, and the responsibility for R&D concerning energy efficiency does not reside with the utilities. It is assigned to the Public Interest Energy Research (PIER) group at the CEC. PIER is primarily engaged in technology development and demonstration. There do not appear to be any projects within the PIER structure designed to improve existing energy efficiency program delivery systems.

2. There are no reports to the regulators by utilities concerning the results of R&D efforts being undertaken in conjunction with the development or improvement of energy efficiency programs.

3. Until very recently, there has been no specific discussion of R&D activities related to improving program performance on the part of utilities. Improvements in program performance are expected to be derived from information obtained from process evaluations – research that is conducted in the course of full-scale program operations – not in the course of program development.

4. Thus far, CPUC has not authorized or earmarked payments to utilities to engage in R&D to support energy efficiency program development beyond technology demonstration projects and pilots of new program ideas.

This is not to say that programs were developed haphazardly. Their development has followed a path. However, it is a path that usually involves taking a shortcut around an important stage in the process of innovation (i.e., experimentation with alternative product design elements during the product (program) development stage). It is a path that runs directly from the concept testing stage to the test marketing stage (i.e., pilot test), or beyond, to implementation of the full scale program.

This shortcut probably brings potential energy efficiency programs to market sooner than would a process involving carefully designed small-scale experiments intended to test alternative program design features prior to test marketing. It also probably dramatically increases the risk that energy efficiency programs moving from the concept testing stage to the operational stage will be ineffective and difficult to terminate.[4] In such situations, the programs themselves may become experiments – very large and expensive experiments that produce information of questionable value very slowly.

One might imagine that market tests or pilot programs themselves serve the same purposes in the development of energy efficiency programs as the Program Development, Test Marketing and Production phases taken together. There are several reasons why this is not a good way to think about the role of market tests. First, market tests are the last stop on the way to program development. They are not meant to test the efficacy of different program design alternatives. They are meant to test the eventual performance of the fully developed prototype prior to full-scale production. Clearly, it is possible to determine from a market test whether a given program prototype can achieve its objectives at a given time and place. However, the information obtained from a market test generally is not very useful for making detailed product or program design changes unless a number of market tests are carried out while varying program design features. In which case, the market tests are experiments. Depending on the size and duration of the market test, this can be a very expensive approach to testing design alternatives.

To understand the difference between the kinds of experiments that are being advocated in this paper and the kinds of pilots or demonstrations that typically take place in energy efficiency program development, it is useful to consider an example of a program design problem and how it might be approached through experimentation on the one hand and pilot testing or full-scale rollout on the other. The example actually concerns the development of a demand response program, but the lessons to be learned about the usefulness of experimentation are the same for both kinds of programs.

Experimentation Example: Demand Response Program Development

A major west coast utility currently has more than 300,000 residential customer participants signed up for a load control program that allows them to cycle or shed customer air-conditioning load during emergencies. Incentive payments on the program vary with the size of the air conditioner, the degree of cycling (50%, 67%, and 100%) and the frequency with which cycling can occur (a maximum of 15 times or unlimited). For a 3-ton air conditioner, payments can range from as little as $18 per summer to as high as

[4] Once programs are placed within the organizational context, they often develop institutional supporters who may be reluctant to terminate an otherwise unsuccessful effort. Examples of institutional supporters include: program managers, evaluators, senior management, regulators and interveners. These parties may instead encourage incremental changes over funding cycles to try to improve performance instead of recommending that ineffective programs be killed outright.

$130 per summer. The vast majority (89%) of program participants has selected the 100% cycling strategy, and 75% of all customers have elected to allow the utility to cycle their air conditioners an unlimited number of times in a given summer. In other words, they have selected the combination of cycling severity and frequency that produces the maximum possible discount for participation.

The program has been used infrequently and primarily for localized distribution problems rather than across the full participant base. Consequently, few customers have ever been cycled. To improve the economic performance of the program, it could be converted from the current, capacity-based, incentive-driven reliability resource to a performance-based (i.e., customers are paid when they curtail) resource and dispatched more frequently based on economic rather than emergency criteria. However, moving from a capacity-based dispatch model to an economic dispatch model while retaining existing customers poses significant program redesign challenges. Significant attrition from the existing program can be expected to occur under almost any circumstances, but some combinations of program features and marketing strategies are likely to be more effective in retaining customers than others.

It is possible to imagine several potentially viable approaches to transitioning from the current reliability-based incentive design to a design based on performance, such as the following:

1. Existing participants could be offered the choice of continuing on the current program at substantially reduced incentive levels (in line with their true economic value) or signing up for a new program that would compensate them (substantially more) on a per curtailment basis commensurate with the value of the energy avoided during high cost periods.

2. Existing participants might be offered the choice between continuing on their current program at substantially reduced incentive levels or signing up for a combination of Critical Peak Pricing (CPP) and/or Time of Use (TOU) rates.

3. Existing participants might be offered choices within a menu of program operational levels (e.g., frequency, notice, duration, etc.) with varying incentive levels – similar to the way the program works currently but with incentives for reliability-based curtailment reduced significantly to its current worth and most of the customer benefit transferred to performance.

4. Existing customers might be assigned to one of the first two performance-based program designs and would be allowed to opt out to a much reduced capacity-based plan.

There are other important program design issues, but the above options are sufficient to identify significant program design alternatives that should be tested during the process of redesigning the program.

It is not obvious which of the program design and marketing strategies in the demand response program example will yield the best results (high retention and high value load relief). Indeed, the choice among the foregoing options can really only be made after observing how consumers respond to them.

Confronted with the above set of alternative designs, the program designer has two basic choices. They can carry out focus groups and customer surveys with samples of customers asking them about their preferences (which needs to be done in any case) and then select an approach that seems most effective – going directly to the CPUC for approval either of a full-scale transition or for a large-scale pilot. Alternatively, they can experimentally offer these program design alternatives to small representative samples (300-400) of customers and observe their reaction. Based on the results from these experiments, the designer can select one of the alternatives, or they might test again based on the results obtained in the first set of experiments.

The alternative that allows for testing of the various program approaches obviously produces much more concrete information about how consumers would respond to different combinations of program design alternatives. It could be conducted very quickly --- if not within 90 days, within 120 days. Moreover, this approach significantly reduces the economic risk to the utility and society involved in the decision (i.e., stranded capital and damaged emergency demand response program capacity). Given the unknowns and the resulting substantial economic consequences, using experimentation to perfect the transition design is certainly cost-justified and almost certainly the right approach in this situation.

Unfortunately, three probably fatal barriers stand in the way of the simple, small-scale experiment as described. First, it almost certainly will be necessary to gain regulatory approval for offering the different experimental alternatives to customers. This may slow the process of program development down considerably and may lead to a complete stop based on the reaction of regulatory staff. Second, there will be resistance by operating departments in the utility to proliferating non-standard billing arrangements (for the customers assigned to different experimental groups). Finally, there is usually some "institutional resistance" to making dramatic changes to program operations that tends to cause key parties in the design process to favor small incremental changes over dramatic ones. People have staked their professional reputations on the outcome of the program, and these considerations sometimes cloud judgment when it comes time to make significant program changes. Taken together, these practical considerations are significant barriers to a very rational approach to program development. This is the sort of situation that typically causes the jump from the concept testing stage (i.e., focus groups and market surveying) to full-scale program implementation in the current program development environment – that, and the ever-present need to act quickly.

With the exception of some notable market failures (e.g., New Coke and mortgage-backed securities), very few products take the risk of shortcutting the innovation process by going from the concept testing stage to test marketing or beyond. Correspondingly, there is little reason to believe this is a reasonable approach to developing effective

energy efficiency and demand response programs. Nevertheless, it happens almost constantly with energy efficiency program development.

It is, of course, reasonable to ask: would an organized product innovation process of the kind conventionally used by industry to develop products and services really improve the effectiveness of energy efficiency programs in increasing the likelihood that consumers would adopt more energy efficient technology and practices? The answer is: it might.

Conceptually, the process of developing energy efficiency programs is the same as the process of developing any other kind of product or service. Moreover, there is good reason to believe that the program development work that is short-circuited in the present process (program design and testing) is precisely the work that needs to be done to discover more effective approaches to improving the likelihood that consumers will adopt more efficient technologies and practices. Finally, as indicated in Section 1, there are a number of potential adjustments that could be made to the existing generation of programs that could significantly improve the performance of existing program delivery systems. So, there is ample reason to believe that innovation management could significantly improve the effectiveness of energy efficiency programs. It remains only to be demonstrated.

3. The Elements of Experimentation

This paper argues there is a significant opportunity to improve the effectiveness of energy efficiency programs by fostering innovation in program design – particularly, in taking advantage of knowledge from the behavioral sciences to develop more effective messages and message delivery mechanisms. As explained above, experimentation is critical to successful innovation. In this section, the kinds of experiments needed to foster innovation in energy efficiency program development are described. This paper strongly advocates the use of scientific experiments as an integral part of the process of innovation in energy efficiency program design. This section dwells heavily on the logic underlying scientific experimentation while providing examples of experimental designs that are particularly useful in energy efficiency program development. Those who are thoroughly familiar with these topics may wish to read only as far in this section as they find the subject matter interesting and useful; and then move on to the conclusions and recommendations section where improvements to the existing program management and regulatory framework are discussed.

At the outset, it is important to distinguish clearly between ad hoc demonstration projects and pilot programs on the one hand and scientific experiments on the other. It isn't that ad hoc demonstrations and pilots are inappropriate to support some aspects of program development. It is that they are not usually designed as experiments and, as such, not useful for sorting through possible program design alternatives to conclusively determine what really works and what doesn't; and this is a critical requirement for fostering innovation in the development of program alternatives in the future.

So far, the use of experiments in program development has been discussed without carefully considering what experiments are and what they are not. Humans use

experiments all the time to solve problems in their daily lives. If we turn on the light switch in our homes and the light doesn't come on (and other appliance in the home are working), we change out the bulb and turn the light switch on again to see if that solves the problem. If that doesn't work, we go to the circuit breaker or fuse panel and reset the breaker or fuse. If the light still doesn't come on, we may try another bulb or maybe change out the light switch. If that doesn't work, we probably call the electrician. In an important sense, these are all experiments. They have in common the fact that we are varying some condition of the world to see if the situation changes.

Some of the experiments that are performed in product innovation are as simple as the one outlined above. This is particularly true when the development of a prototype involves the application of physical processes such as temperature, chemistry, force, etc. However, experiments designed to alter human behavior usually are not so simple. Human behavior, while highly predictable under some circumstances, is inherently more unpredictable than the behavior of physical materials, and this complicates the design of experiments involving humans.

This complication necessitates the use of "statistical experimentation." What distinguishes statistical experiments from the simple experiments done in daily life is that statistical experiments are designed to more or less definitively determine whether a given antecedent condition or set of such conditions causes some characteristic of a population of interest to change and by how much. It turns out that this is really not a very easy thing to do. In fact, there is a very large amount of academic literature governing the design of scientific experiments involving humans.

The discussion in this paper is meant only to introduce the reader sufficiently to the key issues in the design of experiments, so that the utility of such designs in identifying effective improvements in energy efficiency programs can be demonstrated. Readers interested in knowing more about the designs discussed in this paper should read Experimental and Quasi-Experimental Designs for Generalized Causal Inference (Shadish, Cook and Campbell 2002). This book is an excellent and very readable introduction to the topic.

In the 19th Century, John Stuart Mill proposed a set of conditions that must be met in order to show that some condition in the world causes some other condition in the world to change:

1. The supposed cause has to precede the supposed effect in time;

2. The supposed cause must be correlated with the effect – that is, when the cause is present the effect is present, and when it is not, the effect is not present; and

3. No other plausible explanations can be found for the effect other than the cause.

These basic conditions provide the underlying logical basis for the design of scientific experiments.

Moving from the abstract world of philosophy to the world of empirical science, it is extremely rare to find any variable that is caused by a single other variable. Most empirical effects are caused by multiple conditions, and it takes a particular combination of these causal factors to bring about an hypothesized effect. We know, for example, that forest fires can start in a variety of ways – a carelessly discarded cigarette, a spark from a machine, a lightning strike or a smoldering campfire can all start a forest fire. For any of these to actually cause a forest fire, a number of other conditions need to be present. The forest needs to be sufficiently dry, the temperature needs to be high enough, the wind needs to be blowing from the right direction, etc. In practice, we usually don't know how all the potential causal factors interrelate with one another to actually bring about a given effect when we are trying to isolate the effects one of them. So, it is with changing the behavior of humans. Consequently, cause is something most scientists think about probabilistically rather than deterministically. That is, cause is spoken of as something that changes the probability that something will happen.

Before discussing experimental designs that will be useful in discovering effective improvements to energy efficiency programs, it is necessary to spend a few paragraphs describing the overall logical framework that underlies most experimental designs. The fundamental purpose of scientific experiments is to create conditions under which it is possible to correctly infer that changing some condition in the world will cause a difference in the way the world operates. There are certain very important logical requirements that must be met in order to correctly conclude that something has actually caused some other thing. The two most important logical requirements are that the experiment be internally valid and that it be externally valid.

3.1. Threats to Internal Validity

A number of things can happen during an experiment that can cause an experimenter to incorrectly conclude that changing a given condition or set of conditions has changed the probability that some outcome has occurred. In a simple comparison of the outcome of interest before and after exposure to an experimental variable (sometimes called a pre-test/post-test design), there are a number of possible alternative explanations for differences that might be observed besides the operation of the so-called causal variable. The following important possibilities are:

1. History – when we observe a difference in the world at two points in time, it is quite possible that some other factor may have changed in addition to the experimental variable and that this other variable is principally responsible for the observed effect.

2. Maturation – when we observe a difference in the world at two points in time, whether we are observing animate or inanimate objects, it is possible that the object in question matures (i.e., gets older) and something about the aging process causes the change in the outcome measure of interest.

3. Testing – when we observe a difference in the world at two points in time, it is possible that the measurement procedures we are using actually are altering the

situation. Again, looking from one time to another, it may be that our measurements taken in the first time period caused a change observed in the second time interval. This happens all the time in tests with humans when they are tested multiple times using the same procedures.

4. Instrumentation – when we observe a difference in the world at two points in time, it is possible that the calibration of the instrumentation used to measure the outcome of interest changes between the two points in time during which the experiment takes place. Thus, the changes in the outcome measure of interest are due to changes in instrumentation, not to an experimental variable.

5. Statistical regression – when we observe a difference in the world at two points in time, depending on how observations were selected for testing, it may be the case that measurements taken in a second time period are different and closer to the statistical mean of the overall population. This difference can cause us to believe that an effect occurred as a result of the treatment or it can cause the effect to be masked.

6. Mortality – mortality is like maturation except the observed effect of the experimental condition arises from the fact that some subset of a group of observations being taken is not observable at the second time period for reasons unrelated to the experimental condition.

The above problems can be eliminated by comparing what happens to two different groups (as opposed to looking at the same group at two different points in time). The drawback with this approach is that the groups may not have been exactly the same to begin with. This is called:

7. Selection – the groups for which the comparison is being made (experimental vs. control) may have been different before the measurement was taken. So, in this case, there is no basis to infer that the treatment was responsible for all of the differences observed after exposure to the treatment.

All of the above are what has been described as threats to internal validity. That is, they are plausible alternative explanations for why one might observe a difference at two points in time (before and after exposure to an experimental condition) for a given group; or a difference between two groups exposed to different experimental conditions observed at the same point in time. Establishing experimental procedures that ensure internal validity is a critical requirement in experimentation. Experiments that are not internally valid are generally not very useful because they are inconclusive.

3.2. Threats to External Validity

The external validity of an experiment refers to whether or not the results obtained in an experiment can be generalized from the circumstances of the experiment to a broader set of circumstances. That is, whether or not the causal relationships found in the experiment

apply when the persons, settings, treatments or outcomes are changed from the exact conditions observed in the experiment.

1. If the persons or objects observed in an experiment are significantly different from those for which the generalization is to be made, there is reason to suspect that the causal relationship observed in the experiment may not hold.

2. Likewise, it is possible that the experimental treatment works differently in different settings, so that if the setting to which the generalization is to be made is very different from the setting in which the experiment was conducted, there is a possibility that the causal relationship will not hold.

3. If the treatment or outcome measures are changed significantly, there is reason to doubt whether the causal relationship observed during the experiment will hold.

It is possible to overcome the first and second threats (differences in persons and settings) by selecting random samples from the relevant populations of interest (e.g., persons and settings) and assigning them randomly into the experimental conditions.

Controlling the third threat to external validity poses a significant challenge in applied research – particularly, applied research involving outcomes that are to be produced by large organizations. It is possible to create a reasonable small-scale simulation of a marketing process and conduct it with randomly chosen customers to observe the impacts of the process on the likelihood they will adopt the choice that they are given. However, scaling up the experimental prototype to the larger marketing organization can result in changes that cause the actual program operations to be different from what was accomplished in the experiment. As much as possible, to preserve external validity, it is necessary for the actual program to be as similar to the actual treatment as possible. This argues for carrying out field experiments that are as similar as possible to the conditions that will be used in an actual program. On the other hand, integration of R&D into normal business operations is often very difficult to do and can greatly increase the cost and time involved in carrying out an experiment. The loss of experimental control that results may also degrade internal validity. Given these considerations, it is necessary to carefully balance the risks arising from both design alternatives. In the end, it is probably preferable to isolate the organization itself from the experimental process during R&D. Then, if the program doesn't work for some reason, it is possible to isolate the sources of problems in the delivery mechanism.

Most of the literature and thinking concerning experimental design that has evolved over the past 100 years has focused on techniques for ensuring the internal validity of experiments – that is, ensuring that the causal mechanisms that are being described are actually producing the results obtained in the experiment. The discussion of the design of experiments begins with a description of what is generally thought of as the Classic experimental design – the randomized experiment.

3.3. Classic Experimental Design

In the discussions that follow, it is useful to talk in concrete terms about how the various experimental designs can be used to answer important questions about the efficacy of energy efficiency program modifications. Ideally, this section would provide examples of actual experiments conducted in the context of energy efficiency program development to illustrate how experiments are used and useful in the process of program development. Unfortunately, there is very little evidence of the use of experimentation in the development of energy efficiency programs and services in the recent history of the development of these programs in California.[5] Instead, examples will be given of how experimental designs can be used to develop one of the promising new techniques under development at present for improving the performance of mass-marketed energy efficiency programs – statistical targeting.

3.3.1. Developing Effective Statistical Targeting – an Example

For the past several years, statistical targeting algorithms have been under development by a number of parties for improving the likelihood that consumers will respond when contacted. These approaches are in their infancy and there is much to learn about their effectiveness and to improve upon them going forward. For simplicity's sake, this paper will concentrate on the application of experimentation to improving targeting techniques. This thread of innovation is simple to understand and easily lends itself to

---

[5] Techniques from experimental design are sometimes used to evaluate the magnitude of energy savings that were obtained by energy efficiency programs. Pre-test post test designs (a kind of quasi-experiment) are often used to calculate the difference between energy use before and after the installation of equipment to estimate program impacts. Statistical comparison groups (non-equivalent control groups) are recommended in the California Evaluation Framework (CEF) whenever possible as an improvement over the common one group pre-test post-test design (TechMarket Works 2004). However, these commonly accepted quasi-experimental techniques used to assess the performance of energy efficiency programs are not really experiments in the sense that they are meant in this paper. Their purpose is not to discover whether a program design element causes significant improvement or degradation. It is to verify savings. Hence, these studies are more like audits than experiments. There are also references in the CEF to the use of true experimental designs as a means to improve estimates of energy savings – particularly as regards to the problem of understanding the impacts of free-ridership. However, it does not appear that any of the evaluations conducted to date have employed these techniques (TechMarket Works 2004). In the literature on the impacts of pricing on consumption, there are good examples of experiments designed to measure the responses of consumers to demand response initiatives (e.g., dynamic pricing and load management). The best example of this work is the Statewide Pricing Pilot (Charles River Associates 2005; George et al. 2006). Nevertheless, and most importantly, no recent examples were found of experiments designed to test the various alternatives to the design of energy efficiency program delivery mechanisms, such as message content, advertising, targeting, channel effects, social network affects, or other actions that might improve the likelihood that consumers adopt the target technologies and behaviors. While significant efforts are under way to find and demonstrate the efficacy of new energy-efficient technologies (e.g., higher efficiency lighting), there has been almost no systematic effort to find and 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.

experimentation. It should be obvious that the conclusions that are being reached regarding the benefits of experiments for understanding the effects of targeting algorithms apply to efforts to incorporate other behavioral science techniques and theories into energy efficiency program improvements.

The idea behind statistical targeting is simple. Depending on the marketing channels, contact protocols and messages being used, energy efficiency program implementers sometimes churn through thousands of customer contacts to find a relatively small percentage (2-5%) of customers that adopt the energy efficiency actions that are being sold. The basic idea behind statistical targeting methodologies is that information obtainable from the utility and in the public domain (about customers) can be used to identify customers whose propensity for adopting an energy efficiency alternative is higher than that of the average customer. By focusing on these high value targets, the cost-effectiveness of marketing can be improved, and the market penetration of desired energy efficiency alternatives can be increased more quickly.

In theory, targeting could dramatically improve the efficiency of program offerings. However, algorithms for scoring customers for their propensity to respond to energy efficiency programs are in their infancy, although there are multiple technical approaches being offered in the market.

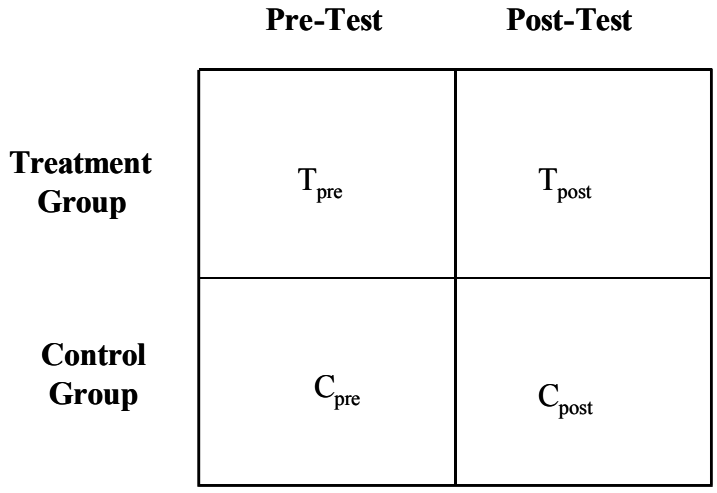
The potential efficacy of improved targeting is more than a theoretical possibility. In a recently reported demonstration of a commercially available targeting algorithm, developers reported that statistical targeting resulted in a five-fold improvement in the likelihood of energy efficiency measure adoption (Willis 2009). Such a large increase in the response rate could dramatically improve the efficiency of marketing. Unfortunately, the reported improvement is based on a demonstration, not on an experiment. What is the difference? The demonstration consisted of comparing the rate of acceptance of an energy efficiency alternative marketed in two different years by the same implementation contractor. In the first year, the customers in the implementation contractor's cold call list were not selected using the targeting algorithm; in the second year, they were selected on the basis of this algorithm. On the surface, this seems like a fairly robust test. However, it fails to control for most of the significant threats to internal validity discussed above. It is possible that targeting improved the response rate. It is also possible that something else happened between the first and second year that caused the improvement in response (history), or that the implementation contractor who knew about the test intensified their efforts or made other changes to their procedure that changed the response. Indeed, virtually all of the threats to internal validity are plausible alternative explanations for the dramatic improvement in response. In essence, it is impossible to conclusively say whether targeting produced the improvement in response, or whether it was produced by some other factor arising from the way the demonstration was done. The point here is not that the demonstration results are wrong or misleading. It is that they are, at best, inconclusive. To obtain conclusive evidence of the benefits of targeting, a true experiment must be conducted. But what is a true experiment? And how is it different from a demonstration?

3.3.2. Completely Randomized Design

The problems with the demonstration described above are not unusual. They have plagued scientific investigations from the very beginning. Around the turn of the last century, Sir Ronald Fisher proposed a simple experimental design that solved these problems by controlling for virtually all of the threats to internal validity. While it is often impossible to employ this design in practical applications, it is useful to understand how it works, because it is the basis for virtually all classical experimental designs.

The most elementary experimental design is called a Completely Randomized Design. It is possible to visualize this design as a four-fold table as indicated in Figure 1.

Figure 1: Block Diagram of Completely Randomized Experimental Design



In this design, observations are randomly assigned to treatment and control groups. The treatment group is exposed to the experimental factor. The control group, which is statistically identical to the treatment group, isn't exposed to the experimental factor. Random assignment effectively eliminates the possibility of selection effects; that is, the possibility that the groups were somehow different at the outset of the experiment. The use of the control group eliminates all of the other possible alternative explanations for an observed difference between the groups because the control group experiences the same history as the treatment group, matures at the same rate, is exposed to the same measurement protocols, and experiences the same mortality. Both the treatment and control groups of observations are measured on the outcome variable of interest before and after the experimental factor is introduced. Observations in this simple experimental design can be anything that can be subjected to observation (e.g., pea patches, people, buildings, businesses, etc.).

The effect of the experimental variable is measured as the difference between differences.[6] That is:

$$\text{Effect} = (T_{post} - T_{pre}) - (C_{post} - C_{pre})$$

This very simple design provides an unambiguous measurement of the effect of the experimental variable on the outcome variable of interest that can be readily subjected to statistical tests for purposes of determining whether the observed difference could have occurred by chance alone given the sizes of the samples involved. It rests on the assumption that the experimenter has complete control over the composition of the experimental and control groups and the presentation of the treatment variable. For reasons that will be discussed below, this situation is difficult to achieve in applied research settings.

It is possible to imagine a simple, yet powerful experimental test of the efficacy of a given targeting algorithm using a completely randomized design. This test involves first randomly selecting a group of 1,000 customers from the total population of customers who are eligible for the program offer. Next, customers are randomly assigned into the control and treatment conditions. The customers in the treatment group are scored according to the targeting algorithm, and they are contacted in declining order of the a priori estimated propensity to participate (from the targeting algorithm). The control group is contacted in random order. Exactly the same marketing should be undertaken for both groups except for the order in which they are contacted. The organization carrying out the customer contact should not be aware of the test.

The effect of the scoring algorithm can be observed by comparing the average value of the order variable (i.e., where the target respondent appears in the call list) for the treatment and control conditions. If there is no effect of the targeting algorithm, there will be no difference in the average order values for successful marketing contacts in the two groups. On the other hand, if the propensity score is highly correlated with likelihood of acceptance of the offer, the average order value for customers who accept the offer in the treatment group should be significantly lower than it is for the control

---

[6] In practice, it is often impossible to obtain pre-treatment measurements for a variety of practical reasons. An experiment designed to assess the impact of a given targeting algorithm on program response rates is such a case. The outcome variable of interest is the change in the likelihood that prospective energy efficiency program participants adopt the technology or practice that is being sold based on some sort of propensity score. There is really no way to observe the likelihood that a given participant will take the subject offer without offering it to them. So, it is impossible to obtain pre-treatment measurement on either the treatment group or the control group. However, the absence of a pre-treatment measurement is not really a problem provided the sample sizes in the experiment are large enough so that the standard error of the outcome measurement is small enough to detect the size of difference that is considered meaningful from a practical standpoint. This is so because random sampling guarantees that the expected values of the outcome measure (i.e., the likelihood or average) are equal for the treatment and control conditions to within plus or minus a known statistical error rate.

group – because the high propensity customers are concentrated at the beginning of the treatment list. Further, using log linear modeling techniques it will be possible to estimate how much impact the propensity score variable has on the actual likelihood of participation – thereby making it possible to quantify the potential magnitude of the efficiency improvement. Unlike the demonstration cited in the example of targeting, the results of this experiment are capable of being logically conclusive. The treatment and control groups are statistically identical to each other, so any difference in the order variable has to be the result of targeting. Of course, if the difference is not large or the treatment and control group sample sizes are too small, the results may be statistically inconclusive.

The usefulness of the Classical Experimental design in practical applications is limited by the fact that many interesting empirical effects (e.g., change in the likelihood that consumers adopt energy efficient technology) arise from the operation of multiple causal factors and while the process of randomization ensures these factors do not systematically affect the experimental outcome, they are capable of producing substantial noise or random variation in the outcome variable of interest. This can cause the operation of the causal variable of interest to be muted by or masked altogether in the outcome of an experiment. This problem has led researchers to elaborate on the randomized design in several ways. They are discussed below.

3.3.3. Randomized Blocks Design

As discussed above, there is bound to be a certain amount of noise in the measurement of the effect of an experimental variable. For example, it is possible that a targeting algorithm is effective under some circumstances but not others. It might work for small, stand-alone, owner-occupied grocery stores but not for chain stores or big box retailers. Or it may work for some kinds of energy efficiency program offers, but not others. There are all kinds of possible confounding factors. This is an example of experimental noise.

It is possible to control for experimental noise by carrying out the randomized design for blocks of customers stratified according to the variable that is suspected of producing it. This is called a randomized blocks design, and it basically involves repeating the completely randomized experiment for customers within the different blocks or strata. Figure 2 displays a Block Diagram of this experimental design. It is nothing more than a series of four-fold tables like the one in Figure 1 – one for each of the blocks or strata. The benefit of this design is that it is possible to detect the effect of the treatment variable within the blocks and between them – so it is possible to describe what is called the main effect of the treatment variable within the blocks as well as the effect of the blocking variable to the extent that is of interest. This approach to experimentation is analogous to stratified random sampling in surveying.

Figure 2: Block Diagram of Randomized Blocks Design

Blocking Factor	Group	Pre-Test	Post-Test
Stratum 1	**Treatment Group**	T_{pre}	T_{post}
	Control Group	C_{pre}	C_{post}
Stratum 2	**Treatment Group**	T_{pre}	T_{post}
	Control Group	C_{pre}	C_{post}
Stratum 3	**Treatment Group**	T_{pre}	T_{post}
	Control Group	C_{pre}	C_{post}

The main effect of the experimental variable is measured as the weighted (by the number of observations) average of the experimental effects observed in each of the strata.

Let's return to the targeting example to see how a Randomized Blocks Design could be useful in measuring the improvement in program performance that arises from statistical targeting. The basic idea behind targeting is that the advance knowledge of important customer characteristics can improve the likelihood of contacting customers that will act on an energy efficiency program offer. In a perfect world, the targeting algorithm would be designed to take account of all of the customer characteristics that could be known about in advance of making an offer that could influence the outcome of an offer. In this perfect world, there should be little to be gained from blocking – because the scoring algorithm should take account of all the differences in customer characteristics that should be controlled to reduce noise in the experiment. This is an important point to keep in mind about blocking. It conveys no benefit and even can degrade the statistical reliability of an experiment if the blocking factor has no effect on the dependent variable.

Of course, the real world is far from perfect, and there may be a number of customer characteristics that influence the likelihood of participation that simply aren't included in

the targeting algorithm, either because the algorithm is too simplistic or because advance information about these factors is not available. For the moment, however, let's assume that the targeting algorithm takes account of virtually all of the customer characteristics that might influence the outcome. Is there no need for blocking?

Actually, even in the situation where the targeting algorithm takes account of all the customer characteristics that might influence the decision to adopt an energy efficiency measure, there are still important controllable sources of variation that may mask the effect of the targeting algorithm in a completely randomized experiment. One that immediately comes to mind that can strongly affect the likelihood consumers accept a sales offer is the unique effect of the sales person/team. Sales persons/teams sometimes vary tremendously in their effectiveness and productivity as a result of differences in experience, training and motivation. This is essentially a random variable that could strongly influence the outcome of a sales offer. In any test of the effectiveness of marketing, this variable is a very good candidate for control, both from the point of view of isolating noise and from the point of view of gauging the eventual effectiveness of the targeting algorithm.

The effects of this variable are relatively easy to control in a Randomized Blocks Design. In such an experiment, the sales person/team becomes a block. If there are two sales persons/teams, there are two blocks. If there are four sales persons/teams there are four blocks, and so on. Each block is given a randomly chosen set of treatment (ordered by targeting score) and control cases. In essence, the completely randomized experiment is repeated for each block. Depending on the number of blocks involved in the experiment, it may be necessary to increase the overall sample sizes to obtain minimal statistical precision in all the cells.

The benefits of blocking in this manner are two-fold. First, blocking on this variable will remove a potentially large source of random variation from the measurement of the impact of the targeting algorithm, thus allowing for a more precise estimate of the unique effect of the targeting algorithm. Second, it will allow for meaningful quantification of the variation in the effectiveness of targeting controlling for the impact of the marketing team – an important source of uncertainty about the likely future effects of targeting in another year or at another location.

3.3.4. Factorial Designs

Not all experimental variables that one might wish to manipulate are binary (present or not). Some experimental variables have levels that are hypothesized to produce different effects on the dependent variable. One example of an independent variable that immediately comes to mind for experiments related to energy efficiency is the incentive level associated with the offering. The impact of incentives on consumer behavior is complicated. In some cases, if incentives are not large enough, no behavior change will occur. However, it has also been shown in some cases that once a certain threshold incentive level is reached additional incentives do not improve the likelihood of the behavior of interest. So, it is not just the presence or absence of incentives that is likely

to produce a behavioral change, but the magnitude of the incentive. These kinds of experimental effects can be measured in what are called factorial experiments.

In a factorial experiment, an experimental variable that can take on different levels is called a factor. The factor is presented in the experiment at different levels, and the objective is to measure the change in the experimental outcome caused by the change in the level of the factor.

So far, discussion has centered on experiments designed to detect the effects of a single targeting algorithm – either present or absent. To see how factorial designs might be useful in assessing the impacts of targeting, it is necessary to consider a situation where there is more than one level of the targeting algorithm. Instead of a single algorithm, imagine a situation where there are two possible targeting algorithms. One algorithm is the information that is readily obtainable by the utility from its own records and can be easily updated internally at low cost (e.g., utility bill records, prior adoption of utility sponsored energy efficiency improvement, SIC code, etc.). In addition to this simple and inexpensive approach, there is another targeting algorithm offered by a commercial vendor that is driven both by the information available to the utility and the proprietary information that the utility must purchase on a subscriber basis (e.g., years in business at location, ownership status, number of employees, revenues, etc.). The question is: how much does the higher cost targeting algorithm improve the likelihood that consumers adopt the offered energy efficiency product or service? A factorial experiment would be useful in this situation for observing the effects of the different levels of targeting.

The two different targeting algorithms represent different levels – one simple and relatively inexpensive and the other more complex and relatively expensive – of a single factor (targeting). In a factorial experiment, observations are randomly assigned to differing levels of the experimental factor as well as to treatment and control conditions. In this case, the factor has two levels – call them low and high cost. For this experiment, a representative sample of 2,000 customers would be drawn from the eligible customer population. Half of this sample would be randomly assigned to the low cost experimental level and half would be assigned to the high cost experimental level. Then, the groups within each level would be randomly divided again into treatment and control conditions.[7] As in the case of the completely randomized experiment, the treatment group in Level 1 would be sorted and processed in declining order of the propensity score from the low cost targeting algorithm. The control group in Level 1 would be processed in random order. The treatment group in Level 2 would be sorted and processed in declining order of the propensity score for the more expensive algorithm. The control group for Level 2 would be processed in random order. The effects of targeting would be measured as described in the discussion of the completely randomized experiment.

[7] It is often possible to suppress repeated control groups and develop an aggregate control group to obtain economic efficiencies. In this case, there is a single control group that provides a standard against which both levels of the factor are observed.

It is also possible to study the effects of more than one factor at a time in a factorial experiment. Figure 3 shows a block diagram of such an experiment

Figure 3: Block Diagram of Factorial Experiment

			Factor 2	
		Group	Level 1	Level 2
Factor 1	Level 1	Treatment Group	T_{11}	T_{12}
		Control Group	C_{11}	C_{12}
	Level 2	Treatment Group	T_{21}	T_{22}
		Control Group	C_{21}	C_{22}
	Level 3	Treatment Group	T_{31}	T_{32}
		Control Group	C_{31}	C_{32}

This is a straightforward extension of the simple factorial experiment, except instead of one factor, there are more than one. For example, another factor that might have an effect on a consumer's decision to adopt an energy efficient technology is the magnitude of the incentives that are being offered in the program. Another possible factor is the marketing strategy that is being used to contact the customers. It might be, for example, that a contact protocol involving a pre-contact letter from the utility explaining that they will be soon be contacted concerning their interest in participating in an important energy efficiency improvement program significantly enhances the likelihood they will listen to the coming cold call eliciting participation. Using a factorial design, it is possible to structure an experiment that is capable of measuring the combined effect of these two variables on consumers' decisions.

The combined effects of two experimental variables can occur in three ways. First, it is sometimes the case that the combination of two factors has a multiplicative effect on a dependent variable. That is, the effect of one of the factors magnifies the effect of the other. This is called an interaction effect. Interactions indicate that the variables working in tandem produce significantly stronger or weaker effects than would be expected if only

one of them was present. In trying to improve on the effectiveness of energy efficiency program delivery systems, this is precisely the sort of relationship that one should be looking for – something that increases the leverage of the aspects of the program that are already in existence.

If the variables do not interact, it is possible to observe two other kinds of effects called the main effects of the factors of interest. Main effects are essentially the unique effects of one of the (in this case) two factors in the experiment. In this case, it would be the effects of the incentives alone or the effects of the information treatment alone on the dependent variable.

While it is possible to imagine testing more than two factors in a single experiment, care has to be taken in the design process to ensure that the interactions of the variables are interpretable. Interactions involving more than two variables are difficult to interpret, and the whole point of doing a factorial experiment is to find interactions that are understandable.

It is possible to form a large variety of hybrids experimental designs using combinations of the foregoing basic ideas. Readers interested in a further discussion of classical experimental designs should consult *Experimental Designs: Second Edition* (Cochran and Cox 1976).

3.3.5. Covariance Designs

The Randomized Blocks design described above can control for a small number uncontrolled factors (e.g., sales person/team) that can influence the outcome variable of interest. However, the effectiveness of this design depends critically on having advance knowledge that the experimental affect varies significantly within values of the blocking variable(s). Blocking on a variable for which this not the true will not reduce the noise in the experiment and will generally lead to lower statistical power.

An alternative approach to blocking that does not depend as much on prior information about the effectiveness of the blocking factor and which allows for a larger number of control variables is called the Covariance Design. In the Covariance Design, the experiment is conducted in exactly the same manner as the randomized experiment. However, in addition to the outcome variable of interest (in this case, the likelihood of responding to the marketing contact), measurements are taken on all of the variables that are thought to influence the likelihood of participation ("covariates"). Examples of possible covariates for the current example could include:

- Sales person/team
- Number of contact attempts
- Customer business type
- Number of employees
- Ownership type
- Building type
- Usage

- Demand
- Likelihood that the business would adopt the energy efficiency measure before the test (if known)
- Education of the decision maker
- Impact of the decision on the customers' bills
- Prior history of adopting energy efficiency measures
- Attitudes and opinions of the management about the impact of energy use on climate change

The variation in the composition of the groups under study with respect to all of the uncontrolled but potentially powerful independent variables will produce noise in the measurement of the effect of the outcome variable. This noise can be greatly diminished by controlling for the correlations among the dependent and uncontrolled independent variables analytically (i.e., after the fact of the experimental manipulation).

In a covariance design, the values of the uncontrolled independent variables are observed (usually) before the experimental treatment has occurred for both treatment and control groups. It is possible, therefore, to estimate a regression function that predicts the mean of the outcome variable of interest from the uncontrolled independent variables for both the treatment and control groups.[8] These regression adjusted means or proportions are then used to estimate the values of the outcome variable of interest in the treatment and control conditions. That is, instead of comparing simple means or proportions for treatment and control groups, we are comparing regression adjusted means or proportions for these groups.

Figure 4 displays a hypothetical example of a covariance design in which the outcome of an experiment (in treatment and control groups) is analyzed by comparing regression adjusted means. To the extent that the variables in the regression functions more or less precisely predict the values of the dependent variable, they will produce much more statistically precise estimates of the dependent variable than the sample means or overall proportions observed in the treatment and control groups without adjustment. Of course, if the predictive power of the regression models is low, the improvement in statistical precision will not be significant. In studying consumer behavior related to energy use, covariance designs are extremely useful.

[8] In practice, it is not necessary to calculate a separate regression equation for treatment and control groups. Instead a single regression equation containing a unique intercept parameter for subjects in the experimental condition and control conditions is usually used. It is necessary in carrying out an analysis of covariance to verify that the values of the uncontrolled independent variables did not somehow interact with the experimental treatment. This should not have occurred because the subjects were randomly assigned to the treatment and control conditions. However, in studies where random assignment was not part of the experimental design, discovery of such interactions is required.

Figure 4: Example Analysis of Covariance Adjusted Means

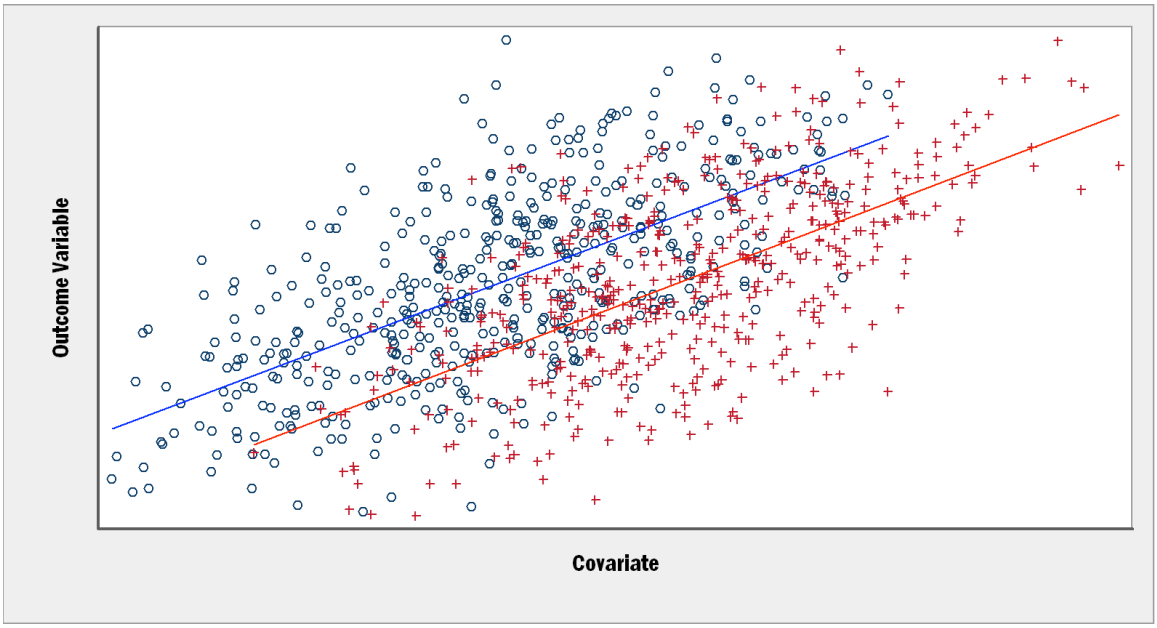


Figure 4 displays a scatter plot of the results of the experiment. The variable on the y axis is the outcome or dependent variable. The variable on the x axis is the covariate. In the figure, it is evident that the dependent variable generally increases with the value of the covariate. However, there is a difference in the effect of the covariate for the treatment (represented by blue circles and blue regression line) and the control condition (represented by the red crosses and red regression line). Inspecting the graph carefully, it is apparent that while the swarm of points is relatively wide, it is generally the case that the blue circles are above the red crosses. The regression lines are parallel (i.e., they have the same slope) indicating the effect of the covariate is the same in both treatment and control conditions. However, the intercepts are different – the intercept of the blue line being above the intercept for the red line. This is the effect of the experimental variable.

The result depicted in Figure 4 is but one of many kinds of effects that might arise in an analysis of covariance. For example, it is possible that the regression lines are not parallel, but instead cross at some point. This is a circumstance where the effect of the covariates varies with the treatment condition. It is what is called an interaction between the treatment condition and the covariate. When this occurs, it is impossible to interpret the main effect of the treatment independent of the effect of the covariate because the difference between the slopes changes as the covariate changes. While such a finding makes the interpretation of the relationship between the covariate and the treatment more difficult, it is no less informative than a simple main effect.

The example in Figure 4 contained only a single covariate. However, the analysis of covariance as described above can be applied to situations in which many covariates are included in the regression equation. In such a situation, particular attention has to be paid to ensuring that covariates do not interact with the treatment, but there are

conventional and widely accepted regression modeling procedures to find and evaluate such interactions.

3.4. Quasi-Experimental Designs

All of the Classical experimental designs described above have in common the fact that observations (e.g., subjects, households, businesses, etc.) are randomly assigned to treatment and control conditions. For many experiments that will be required to assess hypotheses concerning the impacts of different program design options on the likelihood that consumers adopt more energy efficient technology, it will be possible to randomly assign observations to experimental conditions and use the Classical designs. This approach is definitely to be preferred when it can be applied.

However, occasions frequently arise when random assignment to experimental conditions is impossible. There are at least two important classes of R&D problems that fall into this category.

1. It is impossible to use random assignment when exposure to the treatment condition of interest is compulsory, or when observations have the ability to select whether or not they are subjected to the experimental condition. There are well known examples of this situation in medical science. For example, in studies related to the efficacy of medical treatments, it is unethical to withhold treatment from parties who could benefit from it. So, it is often impossible to randomly assign patients to experimental conditions in experiments concerning the efficacy of medical treatments.

   These issues also sometimes arise in studies of energy consumers served by utilities. For example, the purpose of the California Statewide Pricing Pilot (Charles River Associates, 2005; George and Farouqui 2006) was to determine the extent to which consumers would reduce their energy demand and usage in response to time differentiated pricing including a critical peak pricing (CPP) signal.[9] While researchers argued strongly that random samples of customers should be assigned to the treatment and control conditions in this experiment, the fact that publically funded programs must be made available to everyone led to the situation in which consumers were allowed to opt themselves <u>in</u> to the experimental condition. The control group was then selected as a random sample of customers within classes, and differences arising from selection were controlled for analytically (insofar as it was possible to do so).

[9] The Statewide Pricing Pilot was an experiment in which representative samples of residential and small non-residential electric customers in California were assigned to different electric rate designs involving different types of time varying rates (e.g., time of use rates, rates that varied on a daily basis based on the cost of service, and current rates) as well as technologies used to assist customers in responding to pricing changes. The critical peak pricing (CPP) design involves charging a significantly higher price for electricity during afternoon hours on a limited number of days when the cost of electricity production is at its highest level.

2. There is a second important class of experiments (and we use the term very loosely here) in which the experimenter has virtually no control over the assignment of observations to treatment and control groups. A good example is one in which an effort is made to measure the impacts of energy efficiency programs undertaken by organizations or firms. These experiments involve trying to observe the effects of programs offered in schools, or by local, regional or statewide governments.

   Problems of this kind are actually quite common in evaluating the effects of programs in education and public health. In such studies, the organizations that received the treatment are matched on relevant characteristics with those that did not, and the differences between these "control" and treatment communities are attributed to the treatment condition.

As discussed above, when random assignment to treatment conditions is impossible, the design of experiments is a much more complicated problem. When observations are randomly assigned to treatment and control conditions, the plausible alternative explanations (e.g., history, maturation, etc.) for an observed experimental effect are logically and mathematically eliminated. When this is not so, it is necessary to structure the experiment in such as way to observe whether these alternative explanations are plausible, measure their magnitude, and if possible, control for them analytically. This is the domain of quasi-experiments.

The design of quasi-experiments is something of an art, and a large number of such designs have been developed over the past four decades – too large to discuss comprehensively in this paper. Readers interested in Quasi-Experimental designs that are particularly appropriate for guiding the development of energy efficiency programs should turn to Appendix A where several interesting applications are discussed.

3.5. Closing Thoughts on the Use of Experiments to Foster Innovation

This section has attempted to acquaint the reader with a general understanding of experimentation and how it can profitably be used to foster innovation in energy efficiency program development. It is a complicated subject in some ways, and the reader might be wondering: does the problem really have to be this complicated, and does everything have to be subjected to experimentation? The answer is obviously not. Experiments are potentially expensive and time consuming. So they should not be undertaken without considering the benefits that are to be gained and the probable costs in terms of resources and time. The key to success in innovation lies not in slavishly applying experimentation to test all the possible improvements to programs that one can imagine. It lies in strategically applying experimentation to obtain answers to critical questions about what works and what doesn't – particularly when testing the impacts of new and potentially very promising techniques.

4. Conclusions and Recommendations

Innovation in the development of energy efficiency programs isn't really an option, it is a requirement. Ineffectiveness in energy efficiency program marketing is a serious and

largely unnecessary waste of societal resources that are being dedicated to addressing the most important challenge American society has encountered in generations. Innovation depends heavily on the use of small-scale experiments designed to identify program designs that work and those that do not. This paper has briefly outlined the basic elements of experimentation and presented examples of simple experimental designs that can be easily implemented in the context of program development to improve effectiveness. The paper used a single, relatively simple example to illustrate how small-scale experiments can be used to improve program performance. In addition to targeting enhancements, there are many other possible program improvements that should be developed. The effects of different messages, alternative delivery channels, alternative marketing contact protocols, social influence and a wide variety of other factors may be even greater than targeting improvements. Without experimentation, we have no way of knowing.

Experiments cost money, and they require time. How can we be sure that the benefits arising from the use of experimentation exceed the costs in terms of money and time? In theory, they should. However, whether they do or not depends on how intelligently experimentation is applied to program development. One thing is certain. Proceeding along a path in which programs are modified by making small incremental changes from year to year is an inefficient, slow, expensive and largely ineffective approach to program improvement. It is hard to see how using small-scale, carefully chosen experiments to guide program improvement could be worse than the status quo. Like anything else, however, it remains to be demonstrated that the time and effort required to incorporate experimentation into program development is outweighed by the benefits captured in terms of increased program effectiveness – an interesting topic for an experiment.

The need is there. The tools are there. The benefits appear to be potentially large. So, what needs to be done to integrate commonly used innovation management practices including small-scale experimentation into energy efficiency program development. It turns out that this objective is a very tall order. There are several significant barriers to the adoption of the practices described in this paper, as described below.

First, the regulatory process surrounding energy efficiency programs (i.e., the authorization process, funding cycle, evaluation, and oversight functions) in California discourages R&D by utilities designed to improve program performance in advance of full-scale program roll-out:

1. Responsibility for energy efficiency R&D rests primarily with the CEC's PIER organization, not with the utilities. The PIER organization is principally focused on R&D related to new technology, not on program effectiveness.

2. While, impact and process evaluation activities are relatively well-funded in California (i.e., about 8% of program cost) under the heading of Measurement and Evaluation, these activities are not primarily designed to provide information to improve program performance. Instead, these activities are designed to <u>verify</u> whether the claimed savings (or actions taken by the implementers) were realized. The purpose of measurement and evaluation is not R&D. The current focus of

measurement and evaluation is to audit the performance of the programs (i.e., determine whether agreed upon savings or activities have been delivered). Correspondingly, very little information that is useful for improving program performance is generated by this effort. The information is generally too little (because no program design alternatives are tested in the full-scale roll-out) and too late (because it becomes available 2 to 3 years after the program has commenced operation).

3. Except for relatively minor expenditures associated with concept development and testing, there are really no resources set aside in the utility program management and administrative budget for R&D designed to improve program performance.

4. Energy efficiency programs are viewed as energy resources in a broad strategic sense within the regulatory paradigm. Correspondingly, utilities agree to deliver cost-effective energy savings (as opposed to energy) in return for reimbursement for the cost of achieving those savings and financial incentives. Periodically, regulatory bodies and the utilities negotiate energy efficiency savings targets, and the attainment of these targets drives the management of the energy efficiency programs within the utilities. The ability of utilities to recover their costs and obtain incentives depends on achieving savings while minimizing program costs within a given funding cycle. R&D costs that occur within a given funding cycle are applied to the program costs sustained in that cycle. Given that utilities recover the cost of R&D through the cost of administration, significant R&D costs can erode both the ability of the utility to recover its costs and its ability to achieve desired incentives within a given funding cycle. In essence, the current funding mechanism discourages investment in R&D that cannot be recovered within a given cycle. This makes R&D a very unattractive alternative.

Taken together, the above constraints discourage utilities from engaging in significant R&D designed to improve energy efficiency program performance. Unless the constraints are resolved, it is unlikely that utilities will engage in meaningful R&D designed to improve the future performance of their programs. These barriers can be overcome, but it will require an overhaul of the regulatory apparatus that provides funding for R&D that is not tied to near-term energy savings. One way to accomplish this objective would be to build a Chinese wall around the R&D enterprise (within utilities) that encourages them to experiment with interesting program design alternatives and does not penalize them for the normal failures that attend experimentation during a given funding cycle. That is, set aside funding for the R&D enterprise that can be used solely for that purpose and that is explicitly understood to be directed at improving the effectiveness of energy efficiency programs as soon as possible, but not necessarily within a given funding cycle.

Now, this doesn't mean that the utilities should be given a blank check to carry out whatever R&D that they imagine. It means that utilities and regulators have to agree upon reasonable long-term R&D objectives related to the improvement of program performance and jointly prioritize the R&D agenda. It also means that a regulatory

framework for routinely managing the attainment of those goals over time has to be developed.

A second possible impediment to the implementation of significant R&D designed to improve energy efficiency program performance is the lack of experience on the part of both utilities and regulators in the management of innovation:

1. As indicated in Section 2, R&D has to be integrated into business organizations in such a way as to foster innovation without disrupting normal business operations. There are lots of institutional models for how to do this (e.g., centralized R&D department, R&D spread across product lines, etc.). However, it is not obvious what the best organizational structure would be for integrating R&D into a utility energy efficiency enterprise. Since utilities have not been doing R&D of the kind discussed in this paper, it is very likely that their current organization for managing R&D is not ideal from the point of view of fostering innovation.

2. A similar organizational challenge exists on the regulatory side. If a significant R&D is undertaken by utilities to improve energy efficiency program performance, who within the regulatory staff will be responsible for overseeing this effort? Will it be a department, a committee, a public board? Who will staff these entities? What should be their qualifications? What should be their responsibilities? Again, because so little R&D designed to improve performance has been undertaken to date, it is likely that the current organizational structure at the CPUC is not ideally suited to this task.

One might imagine that once the regulatory barriers that are identified above are removed the organizational problems outlined above will resolve themselves organically. That is, that the utilities and the regulatory body will naturally alter their organizational structures and labor forces to take account of the newly arrived R&D goals and funding. That is probably wishful thinking that may result in a lot of unnecessary conflict between the two parties. A better approach would be to instruct both bodies to study the staffing requirements needed to support a significant R&D effort and identify organizational changes necessary to competently implement the proposed R&D program, and to ensure that the funds are being expended prudently.

It is obvious that the above recommended changes to the current operational paradigm for development of energy efficiency programs represent an extremely daunting challenge. It will require significant effort on the part of both the utilities and the regulatory staff to implement the suggested changes; it will be costly; and it will take time to get it right. On the other hand, it is extremely difficult to see how innovation in energy efficiency program development can be fostered without overcoming these difficult problems.

References

Biggart, N. and L. Lutzenhiser, "Economic Sociology and the Social Problem of Energy Inefficiency,"*American Behavioral Scientist* 50(8): 1170-1188 (2007).

Charles River Associates, *Impact Evaluation of the California Statewide Pricing Pilot*, *Final Report*, Oakland, CA, 2005.

Chattopadhyay, P., W. Glick, C. Miller and G. Huber, "Determinants of Executive Beliefs and Company Functional Conditioning and Social Influence," *Strategic Management Journal* 20:763-789 (1999).

Cochran, W. and G. Cox, *Experimental Design: Second Edition*, Wiley Press, NY, 1976.

Darley, J. and J. Beniger, "Diffusion of Energy Conserving Innovations," *Journal of Social Issues* 37(2):150-171 (1981).

Festinger, L., *Theory of Cognitive Dissonance*, Stanford University Press, Palo Alto, CA., 1957.

George, S., A. Farouqui and J. Winfield, "California Statewide Pricing Pilot: Commerical and Industrial Analysis Update," Charles River Associates report prepared for Working Group 3.

Golove, W. and J. Eto, "Market Barriers to Energy Efficiency: A Critical Reappraisal of the Rationale for Public Policies to Promote Energy Efficiency," LBNL #38059, Lawrence Berkeley National Laboratory, Berkeley, CA, 1996.

Kantola, S., G. Syme and N. Campbell, "Cognitive Dissonance and Energy Conservation," *Journal of Applied Psychology* 69(3): 416-421(1984).

Lutzenhiser, L., "Social and Behavioral Aspects of Energy Use," *Annual Review of Energy and the Environment* 18: 247-289 (1993).

Lutzenhiser, L., K. Janda, R. Kunkle, and C. Payne, "Understanding the Response of Commercial and Institutional Organizations to the California Energy Crisis," California Energy Commission, Consultant Report, Sacramento, CA, 2002. .

Lutzenhiser, L., R. Kunkle, J. Woods, and S. Lutzenhiser. 2003. "Conservation Behavior By Residential Consumers During and After the 2000-2001 California Energy Crisis," in *Public Interest Energy Strategies Report*, pp. 154-207, California Energy Commission staff report 100-03-012D, Sacramento, CA 2003 .

Nadel, S. and H. Geller, "Utility DSM: What Have We Learned and Where Are We Going?" *Energy Policy* 24(4): 289-302(1996).

Rogers E. and F. Shoemaker, *Communication of Innovations, a Cross Cultural Approach*, Free Press, NY, 1971.

Rufo, M. and F. Coito, "California's Secret Energy Surplus: the Potential for Energy Efficiency," The Energy Foundation, San Francisco, CA, 2002.

Shadish W., Cook T. and D. Campbell, *Experimental And Quasi-Experimental Designs For Generalized Causal Inference*, Haughton Mifflin Company, 2002.

Sanstad, A. and R. Howarth, "Consumer Rationality and Energy Efficiency," *Proceedings of the 1994 ACEEE Summer Study on Energy Efficiency In Buildings*, pp. 175-183, American Council for an Economic-Efficiency Economy, Washington, DC, 1994.

Sullivan, M., "Behavioral Assumptions Underlying Energy Efficiency Programs for Businesses", White Paper prepared for the California Institute of Energy and Environment and the California Public Utilities Commission's Energy Division, 2009. Available at: http://ciee.ucop.edu/energyeff/behavior.html.

TechMarket Works, *The California Evaluation Framework*, Technical report prepared for The California Public Utilities Commission and the Project Advisory Group, Report No. K2033910. Available at: http://www.tecmarket.net/ca_eval_framework.htm

Thompke, S., *Experimentation matters: unlocking the potential of new technologies for innovation*, Harvard Business School Press, Boston, MA, 2003.

Wenett, R., Kagel, J., Battalios, R. Winkler, R., "Effects of Monetary Rebates, Feedback and Information on Residential Energy Conversation", *Journal of Applied Psychology*, 1978, Vol. 63, No. 1, 73-80

Willis, W., "Using Customer Intelligence to Enhance Energy Efficiency Program Effectiveness," *Proceedings of the Association of Energy Service Professionals' 19th Annual Energy Efficiency Conference and Expo*, January 2009.

Attachment A – Discussion of Quasi-Experimental Designs

A.1 Quasi-Experimental Designs

A.1.1 Non-Equivalent Control Groups Designs

As discussed in the main text, it is sometimes not possible to randomly assign experimental observations to treatment and control groups. When this problem is accompanied by the absence of a pre-test, all of the major threats to internal validity are problematic. In such situations, it is sometimes possible to create non-equivalent control groups whose behavior can be compared with that of the treatment group.[10] These control groups are created by selecting their members from the same population (of people, cities, schools, etc.) from which the treatment group came based on their similarity to members in the treatment group. The idea is to sample people or organizations from the same population that are as similar in important respects as possible to the people or organizations in the treatment group. In essence, it is an effort to "manufacture" a control group that is as similar as possible to the control group that would have arisen from random sampling. This is done by a process called matching.

A number of different matching procedures have been developed including:

1. Exact matching – where each observation in the treatment group is matched exactly with one member of the control group;
2. Caliper matching – where each observation in the treatment group is matched within a range of one member of the control group;
3. Bracketed matching – where each observation in the treatment group is matched with two observations in the control group – one above and one below the score of the treatment observation;
4. Multivariate index matching – where each observation in the treatment group is matched exactly to one observation in the control group based on the value of an index comprising the weighted average of scores on a number of variables; and
5. Propensity score matching – estimates of the probability of selection into the treatment group are used to match members of the population from which the control group is selected with members of the treatment group. This technique requires estimation of the probability of selection using a logit model containing as many known predictors of participation as can be imagined. In simple terms, a logit model is a type of regression model designed to predict the probability that something happens (e.g., participation in a program) based on information about independent

[10] While non-equivalent control groups are sometimes used without a pre-test, this practice is to be avoided because it provides no basis for knowing whether the non-equivalent control group and treatment groups were the same prior to the experimental manipulation.

variables (e.g., education) that are correlated with the occurrence of the event in question. The control and treatment group members are then either exactly matched or stratified into blocks based on their propensity scores, and the results are analyzed in the manner in which Classical experimental designs are analyzed.

Numerous variations on the above procedures have been invented, and their merits have been debated almost endlessly in the literature (Shadish et al. 2002). Matching has a long and dubious history in the experimental design literature. Over time, careful comparisons of the differences between measurements taken from non-equivalent control groups (created through matching) and groups created in randomized designs have revealed that matching comes closest to approximating the results obtained in randomized experiments when treatment and control group members are as similar as possible before matching, and are matched on variables that are statistically reliable and stable. This is a tall order, but possible in some cases. Among the alternative approaches to matching, the consensus is that matching based on propensity scores is the most effective way of matching. Of course, this approach to matching is only possible when useful information is available concerning the variables that may have caused the members of the treatment group to have been selected. To do this, you need detailed measurements of factors that may have caused the selection for parties who were selected into the treatment group and those that were not.

A.1.2 Interrupted Time Series Designs

A time series measurement consists of repeated measures of the dependent variable of interest before and after a treatment has been administered. In energy efficiency and demand response studies, time series measurements are frequently available and extremely useful for evaluating the effects of experimental treatments involving time differentiated pricing (time of use (TOU) rates) and dynamic pricing (critical peak pricing (CPP) and real time pricing (RTP)). The basic idea behind Interrupted Time Series designs is that if the onset time of the treatment is well known it should be possible to observe and quantify a perturbation in the time trend of the outcome variable after the onset of the treatment. This design depends on several important considerations:

1. The onset time of the treatment must be concretely established (i.e., it is definitely known that treatment commenced at a time certain);
2. The effect of the treatment must be large enough to rise above the ambient noise level in the outcome measurement (time series data often contain cycles and random fluctuations that make it difficult to detect subtle effects of treatment variables);
3. If the treatment is expected to gradually impact the outcome of interest, the time series before and after the treatment must be long enough to detect a change in the intercept or slope of the outcome variable after the treatment has occurred;

4. The number of observations in the series must be large enough to employ conventional corrections for autocorrelation if statistical analysis is required (as it almost always is)[11].

Interrupted time series designs are subject to several of the threats to internal validity that accompany experimental designs in general (see Section 3). For example, the observation of a change in the intercept or slope in a time series may have been caused by something other than the experimental factor (History), or it might have been caused by a change in the measuring instrument accompanying the onset of the experimental factor. To control for potential intervening explanations, a variety of quasi-experimental techniques can be employed including: the use of non-equivalent control groups as described above, adding non-equivalent dependent variables (i.e., other variables that are expected to be impacted by the same historical forces as the dependent variable but not the treatment factor); and manipulating the presentation of the treatment factor (adding and removing it) to observe the impact on the outcome variable. The latter is only appropriate when the effect of the treatment factor is expected to be transient.

As indicated above, the Interrupted Time Series design is extremely useful in analyzing the responses of customers to time varying prices and load management signals. It probably would also be very useful in analyzing the behavior of customers in response to almost any kind of feedback concerning their behavior related to energy use that can be observed either in their usage or demand levels.

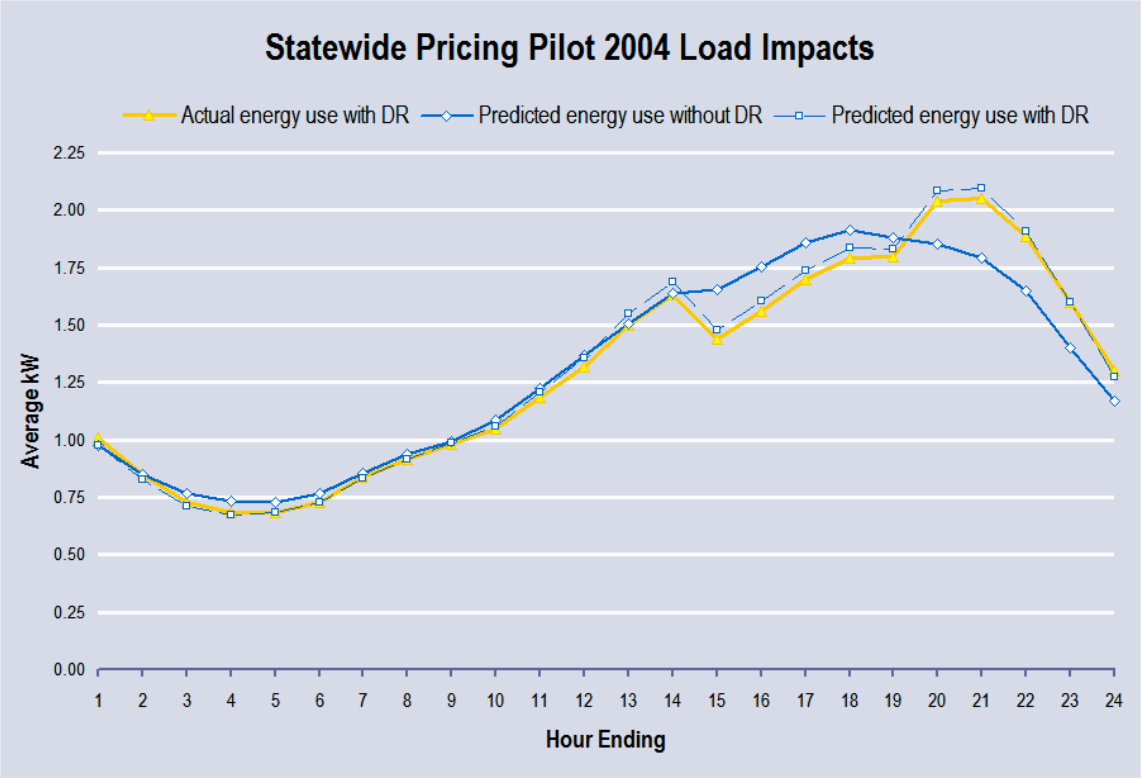
An example of how this technique is applied will illustrate its usefulness. The responses of customers to dynamic pricing signals can be measured at 5 minute to one-hour intervals on a daily basis over the course of a season. With the widespread penetration of advanced metering, this kind of data will soon be available for virtually all kinds of customers. It is possible to send price changes to customers periodically over the course of time. This is the basis of RTP and CPP. It is possible to observe the impacts of changes in prices by observing the changes in customers' loads that correspond with the price changes. To the extent that customers modify their energy use in response to price signals, it is possible to observe this pattern in a time series over the course of the season.

In the parlance of statistics, these designs are referred to as "within subjects repeated measures designs," and they are the state of the art for observing changes in loads and usage in response to price changes. Figure 4 displays the results of a within subjects analysis of the effects of CPP price changes on the loads and energy use of residential customers in the Statewide Pricing Pilot.

---

[11] Autocorrelation is the correlation between measurements of the same variable at different points in time. It is the case that the closer two measurements are to one another in time the more likely they are to be the same. In time series analysis, it is important to correct for autocorrelation when values at a previous time period are used in a prediction model for a value at a later time period. This is called a lagged dependent variable.

Figure A-1: Example of Application of Interrupted Time Series Design



In Figure A-1 the average daily usage on treatment and control days is depicted. That is, the impact of the treatment is inferred by measuring the difference within subjects in the experiment between their hourly electric usage on days when the CPP is in effect and on days when it is not. The figure demonstrates the extent of load reduction that was obtained on the average in the experiment and allows estimation of the net energy savings or gain attributable to the operation of the program.

A.1.3 Regression Discontinuity Designs

It is sometimes the case that the assignment to experimental treatments is governed by strict quantitative criteria. For example, the CPUC has mandated that all commercial customers with annual maximum demand in excess of 200 kW (who are not otherwise participating in a Demand Response program) are required to take service under an applicable CPP TOU rate starting in 2010. They will have 45 days from the start of this requirement to opt out of the rate back into an applicable TOU rate. A number of important questions surround the implementation of this policy, such as:

1. What types of the customers will opt out of the CPP TOU rate option (this is observable once the policy is invoked)?
2. Are structural benefiters (i.e., parties who will automatically benefit from such a rate) disproportionately likely to remain on the rate?
3. How much change will occur in the loads and energy usage of both structural benefiters and others who remain on the rate?

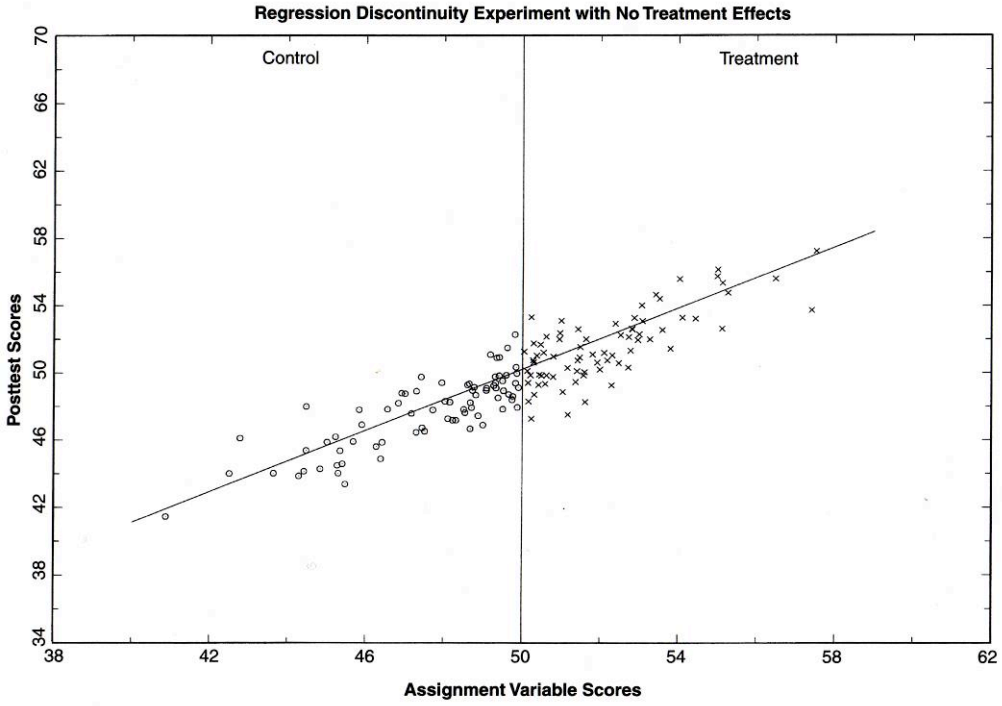
A control group of customers that did not experience the change would be extremely useful in answering the above questions. However, because everyone above 200 kW maximum annual demand will be required to make this change, a control group is not available.

This is a perfect example of an occasion in which it is possible to employ what is called a Regression Discontinuity (RD) design. In an RD design, observations are assigned to treatment control conditions exactly as a function of their score on an interval level quantitative indicator. An interval level indicator is one in which numerical values represent equal intervals of value (e.g., Fahrenheit temperature, altitude, kW demand, kWh, etc.) In an RD design, everyone above or below some point in the scale is assigned to the treatment. So, in this case everyone above 200 kW is assigned to the CPP TOU rate, everyone below is not.

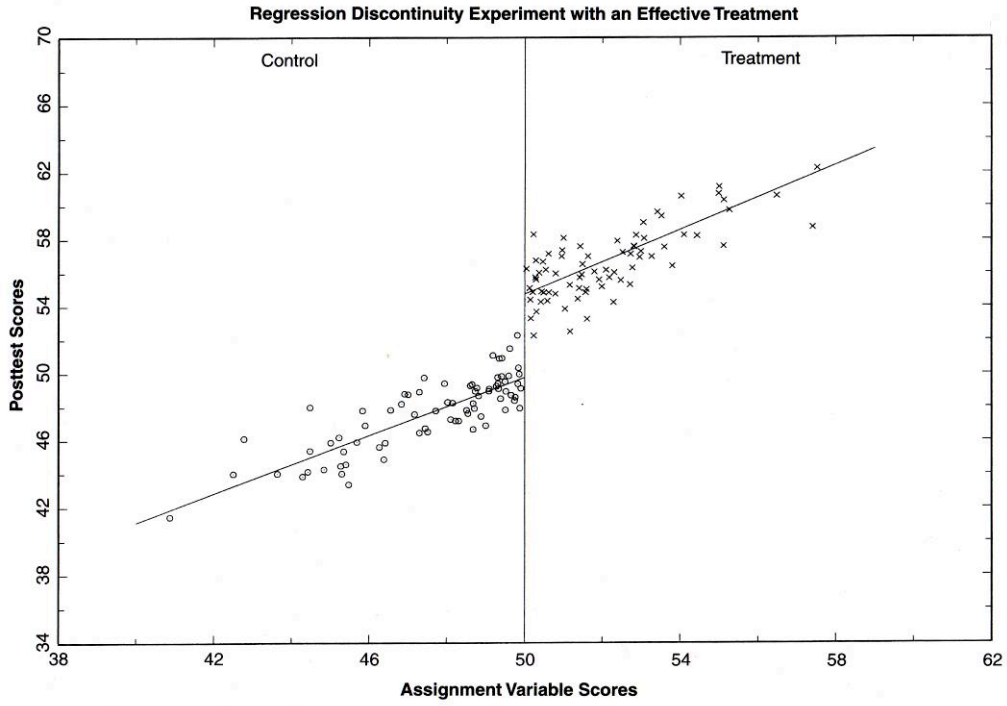
It is possible to specify a regression equation describing the relationship between the assignment variable and the outcome of the experiment. It might be that the outcome measure increases with the value of the assignment variable, decreases with it, or doesn't vary at all. It doesn't matter. What matters is whether there is a difference between the treatment and control groups at the point on the scale variable where the assignment took place.

Figure A-2 displays two examples of the results of an RD analysis. The top panel of the example shows the regression relationship between an assignment variable and treatment outcome when there is no treatment effect. The assignment in this example takes place at the scale value 50. Now, compare the regression relationship in the top panel with the one in the bottom panel. Notice the discontinuity at the point on the assignment scale at which the assignment occurred. The difference in the intercepts for the two regression lines depicted in the bottom panel is the effect of the treatment.

Figure A-2: Examples of Treatment Effects in a Regression Discontinuity Design



Regression discontinuity experiment with no treatment effects.



Regression discontinuity experiment with an effective treatment.

Figure 6 from Shadish, William R., Cook, Thomas D. & Campbell, Donald T., "Experimental and Quasi-Experimental Designs form Causal Inference," 2002, pp. 210-211

This very simple idea is extremely powerful mathematically and statistically. Among all the quasi-experimental designs, this is the only one that is completely equivalent to a Classical experimental design in terms of its internal validity. That is, it controls all of the possible alternative explanations for the observed program effect. There are certain important caveats that must be met:

1. Assignment to the treatment must be strictly determined by the assignment variable. Even the slightest deviation from this requirement will undermine its validity.
2. Care must be taken to remove any crossovers from the analysis (i.e., sometimes parties will migrate into the treatment group from the control group and vice versa).
3. Care must be taken to ensure that the functional form of the regression is correctly specified. If the relationship in the regression is specified as linear and it is not, the regression discontinuity analysis may incorrectly interpret the point of inflection on the non-linear function as a discontinuity, and this will result in a serious error. Likewise, if the treatment interacts with the assignment variable (causing a jackknifed shaped function), and the function is not properly specified as such, this will cause a serious error and one in which the error seriously understates the effect of the experimental treatment.

In this section, only a few very relevant Quasi-Experimental designs have been discussed, and we have barely scratched the surface in terms of the possibilities. The review should have been sufficiently broad to indicate that these designs are applicable to a very wide variety of R&D questions that will arise in the context of finding more effective means of improving the rate of adoption of energy-efficient technologies and practices. The analytical tools are there.