EPRI | ELECTRIC POWER RESEARCH INSTITUTE

Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols

Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols

1020855

Final Report, April 2010

EPRI Project Managers
B.F. Neenan
J.K.G. Robinson

ELECTRIC POWER RESEARCH INSTITUTE
3420 Hillview Avenue, Palo Alto, California 94304-1338 ▪ PO Box 10412, Palo Alto, California 94303-0813 ▪ USA
800.313.3774 ▪ 650.855.2121 ▪ askepri@epri.com ▪ www.epri.com

DISCLAIMER OF WARRANTIES AND LIMITATION OF LIABILITIES

THIS DOCUMENT WAS PREPARED BY THE ORGANIZATION(S) NAMED BELOW AS AN ACCOUNT OF WORK SPONSORED OR COSPONSORED BY THE ELECTRIC POWER RESEARCH INSTITUTE, INC. (EPRI). NEITHER EPRI, ANY MEMBER OF EPRI, ANY COSPONSOR, THE ORGANIZATION(S) BELOW, NOR ANY PERSON ACTING ON BEHALF OF ANY OF THEM:

(A) MAKES ANY WARRANTY OR REPRESENTATION WHATSOEVER, EXPRESS OR IMPLIED, (I) WITH RESPECT TO THE USE OF ANY INFORMATION, APPARATUS, METHOD, PROCESS, OR SIMILAR ITEM DISCLOSED IN THIS DOCUMENT, INCLUDING MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, OR (II) THAT SUCH USE DOES NOT INFRINGE ON OR INTERFERE WITH PRIVATELY OWNED RIGHTS, INCLUDING ANY PARTY'S INTELLECTUAL PROPERTY, OR (III) THAT THIS DOCUMENT IS SUITABLE TO ANY PARTICULAR USER'S CIRCUMSTANCE; OR

(B) ASSUMES RESPONSIBILITY FOR ANY DAMAGES OR OTHER LIABILITY WHATSOEVER (INCLUDING ANY CONSEQUENTIAL DAMAGES, EVEN IF EPRI OR ANY EPRI REPRESENTATIVE HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES) RESULTING FROM YOUR SELECTION OR USE OF THIS DOCUMENT OR ANY INFORMATION, APPARATUS, METHOD, PROCESS, OR SIMILAR ITEM DISCLOSED IN THIS DOCUMENT.

ORGANIZATION(S) THAT PREPARED THIS DOCUMENT

Freeman, Sullivan & Co.

NOTE

For further information about EPRI, call the EPRI Customer Assistance Center at 800.313.3774 or e-mail askepri@epri.com.

CITATIONS

This report was prepared by

Freeman, Sullivan & Co.
101 Montgomery St., 15th Floor
San Francisco, CA 94104

Principal Investigators
M. Sullivan
S. George

This report describes research sponsored by the Electric Power Research Institute (EPRI).

The report is a corporate document that should be cited in the literature in the following manner:

Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols. EPRI, Palo Alto, CA: 2010. 1020855.

PRODUCT DESCRIPTION

Enhanced and timely information regarding consumers' electricity use and costs, known as "feedback," may significantly influence their behavior across a broad spectrum of usage and acquisition decisions. Such information feedback has become increasingly important in light of widespread investments in Smart Grid technologies. This report identifies protocols for three phases of energy information feedback pilots—research design, analysis, and reporting—along with three example applications. When used as the common basis to develop multiple feedback pilot designs, these protocols provide the opportunity to draw meaningful comparisons across pilots, thus helping to increase experimental generalizability and avoid redundancy.

Results and Findings

The 14 protocols presented in this report are widely applicable and largely invariant to the type of feedback experiment being conducted and to the technology used to deliver feedback. The three types of protocols are categorized as follows:

- *Research design protocols* highlight the critical importance of structuring the research design around primary questions and issues of interest. These protocols also provide guidance on population frame definition, sample design, sample size requirements and precision tradeoffs, selection of appropriate control groups (and what to do when control groups cannot be established), key drivers of sampling and control group strategy, and implementation requirements.
- *Analysis protocols* provide guidance concerning statistical methods that are best suited for determining the impact of feedback treatments and for extrapolating results to broader populations (either to target populations within a utility, or to other utilities). The analysis protocols primarily focus on delineating outputs that will facilitate the comparison of results across experiments and allow statisticians and econometricians to assess the quality and validity of the analysis.
- *Documentation protocols* focus on providing a detailed record of the study design and results. In general, these protocols set forth the minimum reporting requirements that should be followed any time results of a feedback experiment are reported.

Challenges and Objective(s)

This report provides guidance in the design and administration of experiments or pilots intended to assess the impacts of feedback mechanisms on consumer behavior. The report has two primary objectives:

- The first objective is to clarify and convey how to design social experiments involving feedback in order to: 1) establish a clear causal relationship between experimental treatments and the outcomes of interest and 2) specify suitable methods for use in analyzing experimental data. This includes extending the research beyond measuring the level of energy usage change to provide a more complete understanding of the underlying behaviors at work.

- The second objective is to suggest methods and output that will allow for more meaningful and robust comparisons across feedback experiments in order to support the pooling of data across experiments. Ensuring comparability across designs is essential in order to avoid redundant research and to help determine whether observed differences across studies are statistically meaningful and, if so, to identify the underlying drivers behind these differences.

Applications, Values, and Use

This report will be of significant value to utility and energy organization personnel planning to design and implement residential feedback pilots, either alone or in combination with other treatment options, such as dynamic pricing tariffs or educational materials.

EPRI Perspective

Research conducted over the past several decades suggests that providing feedback on household-specific energy consumption to consumers can cause a change in the timing and/or magnitude of usage. However, important research questions remain, the answers to which will impact the cost-effectiveness of potential feedback alternatives. Because there are literally billions of dollars at stake in the decisions to purchase feedback technologies and services, it is necessary that such questions be answered conclusively. The protocols in this report will help bring clarity to the process of quantifying the impacts of feedback interventions.

Approach

The authors first identified feedback categories and research gaps. They next defined the elements of experimentation and developed research design, analysis, and documentation protocols. Finally, they synthesized information on design protocol applications and created example applications of the protocols for three different feedback research scenarios.

Keywords

Feedback
Energy consumption information
Consumer behavior
Residential
Consumers
Research protocols
Experimental design
Smart Grid
Conservation
Energy efficiency

ABSTRACT

Enhanced and timely information on electricity use and costs, known as “feedback,” may significantly influence consumer behavior across a broad spectrum of usage and acquisition decisions. This report presents protocols for three phases of energy information feedback pilots: research design, analysis, and reporting. These protocols are widely applicable and invariant to the type of feedback experiment being conducted and to the technology used to deliver feedback.

In all, the report includes 14 specific protocols, with the first 10 addressing research design, the next three covering analysis, and the last one dealing with documentation, as follows:

- Protocol 1—Defining Information Feedback Treatments
- Protocol 2—Determining Outcome Variables to be Measured
- Protocol 3—Delineating Customer Sub-Segments of Interest
- Protocol 4—The Experimental Design
- Protocol 5—The Sampling Plan
- Protocol 6—The Recruitment Strategy
- Protocol 7—Length of Experiment
- Protocol 8—Data Requirements and Collection Methods
- Protocol 9—Minimum Data Requirements for Cross-Utility Comparisons and Pooling
- Protocol 10—Key Support Systems and Materials
- Protocol 11—Load Impact Analysis
- Protocol 12—Behavioral Change Analysis
- Protocol 13—Analysis of Participant Use of Information Feedback
- Protocol 14—Documentation of Feedback Experiments

The report also provides example applications of these protocols for three different feedback research scenarios. The first example application examines feedback via the provision of monthly or quarterly information reports comparing each resident’s energy use with that of neighboring households. The second example application presents a research plan for an experiment involving provision of real-time feedback via different in-home displays employing a PC-based treatment. The third example application focuses on the provision of real-time feedback disaggregated down to the appliance level.

Given the extensive protocols and example applications provided, this report will be of value to utilities and energy organizations planning to design and implement residential feedback pilots, either alone or in combination with other treatment options.

ACKNOWLEDGMENTS

EPRI would like to acknowledge the following reviewers for providing detailed and constructive comments regarding the report content.

- Hunt Allcott, Assistant Professor of Economics, New York University; Energy and Society Fellow, Massachusetts Institute of Technology
- Karen Ehrhardt-Martinez, Environmental Sociologist and Founder, Human Dimensions Research Associates
- Chuck Goldman, Staff Scientist/Group Leader, Lawrence Berkley National Laboratory
- Todd Rogers, Executive Director, Analyst Institute

CONTENTS

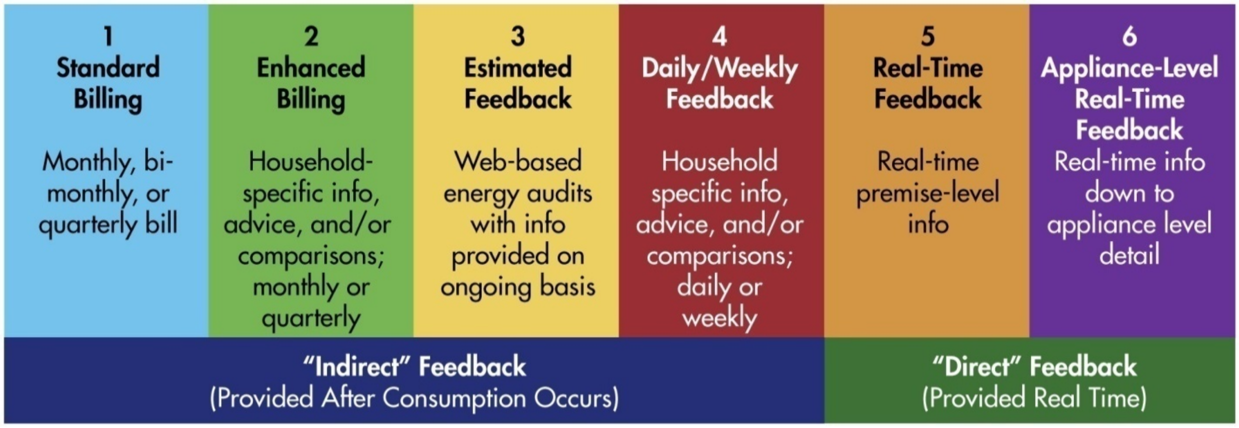
LIST OF FIGURES

LIST OF TABLES

1
INTRODUCTION

The rapidly increasing and widespread investments in Smart Grid technologies occurring in the electricity industry are opening up numerous opportunities for utilities to provide customers with enhanced and timely information on electricity use and costs. The results of a variety of studies suggest that information feedback – "a process whereby the results of action serve continually to modify further action"[1] – may significantly influence consumer behavior across a broad spectrum of purchase and usage decisions.

A recent report commissioned by EPRI[2] reviewed 31 studies and developed a taxonomy of information feedback, an updated version of which is shown in Figure 1-1.



1 Standard Billing	2 Enhanced Billing	3 Estimated Feedback	4 Daily/Weekly Feedback	5 Real-Time Feedback	6 Appliance-Level Real-Time Feedback
Monthly, bi-monthly, or quarterly bill	Household-specific info, advice, and/or comparisons; monthly or quarterly	Web-based energy audits with info provided on ongoing basis	Household specific info, advice, and/or comparisons; daily or weekly	Real-time premise-level info	Real-time info down to appliance level detail
"Indirect" Feedback (Provided After Consumption Occurs)				"Direct" Feedback (Provided Real Time)	

Figure 1-1
Feedback Type Categorization

The EPRI report also identified numerous knowledge gaps in the literature, some of which are being addressed in approximately 30 ongoing studies, most of which focus on Category 5: Real-Time Feedback. Recent awards by the Department of Energy of matching grants for Smart Grid investments will result in – and likely stimulate – numerous additional research studies on information feedback over the next three years.

The various studies currently underway and on the drawing boards have the potential to fill many of these gaps. They also will advance the state of understanding of the relationship between information feedback and consumer behavior. However, without proper guidance concerning research design and analysis and coordination across the various studies, there is the risk that this research will have serious methodological shortcomings and will result in squandering resources,

[1] Webster's Pocket Dictionary, 1997.

[2] *Residential Electricity Use Feedback: A Research Synthesis and Economic Framework.* EPRI, Palo Alto, CA: 2009. 1016844.

duplication of effort, missed opportunities, and misleading findings that can have wide-scale adverse consequences. The purpose of this report is to help reduce those risks so that research and demonstration dollars spent by the electricity industry on information feedback research over the next few years will maximize what we learn about this potential game-changing service enhancement.

This report has two primary objectives:

1. Clarify and convey how to design social experiments involving feedback to establish a clear causal relationship between experimental treatments and the outcomes of interest, and specify suitable methods to be used for analyzing experimental data. This includes extending the research to go beyond measuring the energy usage change to provide a more complete understanding of the underlying behaviors at work.
2. Suggest methods and output that will allow for more meaningful and robust comparisons across feedback experiments to support the pooling of data across experiments. Assuring comparability across designs is essential in order to avoid redundant research and to help determine whether observed differences across studies are statistically meaningful and, if so, what are the underlying drivers of the differences.

Prospective Research Protocols

The discussions that follow focus on how to apply social science methods and procedures to feedback pilots and experiments. However, they are not limited solely to that purpose because they are derived from widely accepted, experimental, and analytical practices.

This report provides guidance to utility analysts and others that are charged with designing and administering experiments or pilots intended to assess the impacts of feedback mechanisms on consumer behavior. The terms “experiments” and “pilots” are used interchangeably here to refer to studies that provide selected consumers with feedback mechanisms under controlled conditions for a specified period, in order to isolate and quantify how feedback influences behaviors. As discussed below, designers are encouraged to go beyond just quantifying the ***impact*** – for example, the reduction in electricity use attributable to the feedback mechanism – and employ research practices that are capable of providing a more comprehensive understanding of how behaviors are modified and how habits are formed.

The experimental framework proposed herein should be distinguished from an ex post statistical analysis which seeks to infer causal relationships from information that is essentially ***historical***, rather than experimental. Using historical data, the conditions under which changes in behavior are observed are not rigorously controlled. As a result, the representativeness of pseudo control groups and statistical adjustments designed to analytically control for differences between those who received the treatment and those who did not are sources of great uncertainty. Biases that can arise in such analyses raise serious questions about assuming, as has been the case until recently, that the results are not confounded with omitted variables (i.e., conform to what randomization would have produced). Rigorous testing is required to first control for potential bias when that is possible.[3] Many ex post experimental procedures, including those that are

[3] The authors are indebted to a reviewer that recommends a valuable source document: M. Bertrand, E. Duflo, and S. Mullainathan. *How Much Should We Trust Differences-in-Differences Estimates?* National Bureau of Economic Research. Working Paper 8841. March 2002. www.nebr.org/papers/w8841

discussed below under quasi-experimental designs that have been widely deployed in the past, may not be appropriate for studying feedback effects after-the fact.

Research Protocols

Guidance concerning how best to meet the specified objectives is provided in the form of protocols. Miriam-Webster's Online Dictionary defines protocol as: "a detailed plan of a scientific or medical experiment, treatment, or procedure." In recent years in the electricity industry, a variety of protocols have been developed to guide evaluations of energy efficiency (EE) and demand response (DR) resources. Some, but not all, are pertinent for the purposes herein.

One class of protocols prescribe the approaches that are to be employed to evaluate programs; for example, California's EE protocols[4] that identify the specific methods that must be applied when estimating load impacts for EE programs in California. These protocols are prescriptive by virtue of their use by utilities for program evaluation as mandated by the California Public Utilities Commission (PUC).

A second type of protocol focuses on the output that must be provided, leaving decisions concerning research methods to be made by the researchers who are responsible for producing the required output. California's load impact protocols for evaluating demand response resources[5] are an example. A third type of protocol primarily provides guidance concerning best practices and recommended approaches to research design and analysis, tailored to a particular subject matter area; for example, demand side management (DSM) evaluation or outage cost estimation. There are numerous examples of these kinds of protocols, usually described as guides or guidebooks, including EPRI's guidebooks for conducting customer value of service studies and for evaluating DSM programs.[6]

The protocols presented herein combine elements of all three types. They are intended to help researchers design productive experiments. There clearly are a number of "right ways" to approach the problem, and these methods and the reasons why they should be used are discussed in this report. They serve to define aspirations in formulating the research design and to establish a benchmark for determining what's lost when design compromises are made. However, there are always intervening circumstances that make challenging to conduct an experiment that conforms to every theoretical dictum. The protocols are intended to provide guidance to experiment designers as they make these decisions. They are worth striving for because the following the protocols described in this report will create opportunities to draw meaningful comparisons by pooling data across studies and testing for differences in outcomes, thereby helping to increase external validity and avoid experimental redundancy.

[4] *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals.* California Public Utilities Commission. April 2006.

[5] *Load Impact Estimation for Demand Response: Protocols and Regulatory Guidance.* California Public Utilities Commission. March 2008.

[6] For example see: *Outage Cost Estimation Guidebook*. EPRI, Palo Alto, CA: 2005. TR-106082 and *DSM Process Evaluation: A Guidebook to Current Practice*. EPRI, Palo Alto, CA: 1992. TR-100647.

Phased Protocols

Protocols are stipulated for three phases of the research process: research design (including establishing research objectives), analysis (ascertaining what was observed), and reporting.

The *research design protocols* begin with planning and highlighting the critical importance of structuring the research design around the primary questions and issues of interest. For example, the research approach and requirements are quite different if the experimental objective is to discover not only *what* happened -- how load and/or behavior changed as a result of an experimental treatment -- but also *why* it happened-- how customers used the feedback associated with a treatment to change their behavior.

They also provide guidance on population frame definition, sample design, sample size requirements and precision tradeoffs, how to select appropriate control groups (and what to do when you cannot do so), the key drivers of sampling and control group strategy, and implementation requirements.

The *analysis protocols* provide guidance concerning the statistical methods that are best suited for determining the impact of information treatments and for extrapolating results to broader populations (either to target populations within a utility, or to other utilities). The analysis protocols primarily focus on delineating outputs that will facilitate an “apples-to-apples” comparison of results across experiments and allow technicians (statisticians, econometricians) to assess the quality and validity of the analysis.

These protocols describe a set of common outputs that include load impacts or summaries of behavioral changes that result from information treatments and selected population characteristics for the treatment group. Data on population characteristics (including weather data) is essential for extending the results to other circumstances. It allows external reviewers and analysts to identify possible reasons for differences in what the utility experiment produced from the results observed in other experiments and/or to assess how different their population is from those who participated in the experiment and, therefore, how applicable the findings would be to their customers. Likewise, the protocols call for providing information concerning the specific historical circumstances in which the study was undertaken, such as unusual environmental, economic, or political events that may have influenced the study results. The analysis protocols also discuss the kinds of validity tests and statistics that should be provided so that a knowledgeable reviewer can assess the quality of the analysis.

The *reporting protocols* focus on documentation of the study design and results, which is closely aligned with the analysis protocols.

The protocols are presented in sections 4 through 6. These protocols are robust, widely applicable and largely invariant to the type of information feedback experiment being conducted and to the technology used to deliver feedback. Of course, the application of the protocols will result in different experiments and analyses because of inherent differences across feedback types and experiments in what is to be measured. Sections 7 through 10 discuss examples of the application of the relevant protocols for three experiments covering information feedback Categories 2, 5, and 6 (as depicted in Figure 1-1). These examples focus on the planning protocols, which determine the nature of the analyses that are subsequently conducted.

Report Organization

The remainder of this report is organized as follows:

- Section 2 briefly discusses the information taxonomy outlined in Figure 1-1 and the research gaps identified in the aforementioned EPRI report.
- Section 3 provides an overview of experimental and quasi-experimental designs.
- Sections 4 through 6 describe the protocols for research design, analysis, and reporting.
- Section 7 provides a brief summary of the reasons for choosing the specific examples contained in Sections 8 through 10 and highlights some of the important differences in objectives and approaches across the three examples.
- Appendix A contains templates containing the protocol tables and worksheets to facilitate their application to research initiatives.

2
FEEDBACK CATEGORIES AND RESEARCH GAPS

The feedback taxonomy depicted in Figure 1-1 serves as the foundation upon which the protocols were developed. As such, it is useful to examine it in more detail, as well as briefly review the body of feedback-related research with regard to overall findings and research gaps.

Feedback Categories

A version of the feedback taxonomy (Figure 1-1) was initially developed by EPRI as a means of comparing the impact evaluation results of different types of feedback programs. Each of the categories is briefly described below, including how they differ from one another.

Category 1 – Standard Billing

- Frequency: monthly or bi-monthly (the baseline scenario for feedback).
- Type of Information Provided: basic information on monthly premise-level kWh and rate ($/kWh), corresponding cost, other fixed charges, and amount due; sometimes includes bar charts showing a comparison historic monthly usage.
- Medium: direct mail (DM) (possibly with a link to a website).
- Information: based on meter reads or estimates.
- Example: typical utility bills.

Category 2 – Enhanced Billing

- Frequency: monthly, bi-monthly, or quarterly, as a supplement to the monthly bill.
- Type of Information Provided: basic premise-level consumption information as in Category 1, but generally including comparative metrics which could be normative (compared to neighbor) or historic (compared to previous consumption) and may include targeted tips (some knowledge of customer and housing/appliance stock necessary); generally more easy-to-read and aesthetically pleasing than standard bills.
- Medium: direct mail or e-mail (possibly with a link to a website).
- Information: based on monthly meter reads or consumption estimates possibly augmented with customer provided information about appliance stock.
- Difference from Category 1: contains targeted information about saving energy, with more detail about household consumption that is likely to be of interest to the customer; the information can be normative, generally more visually pleasing.
- Example: OPOWER "Home Energy Reports," as used by Puget Sound Energy, Sacramento Municipal Utility District.

Category 3 – Estimated Feedback

- Frequency: varies, but information must be provided back to the customer on some sort of on-going basis for it to be considered "feedback" (as opposed to just a one-off "information" web audit program).
- Type of Information Provided: estimated typical premise-level usage, possible estimates of appliance-level consumption, household-specific tips or advice; based on customer-provided data (appliance, house, and billing information, etc.) which is then analyzed to develop estimates.
- Medium: web-based, with ongoing alerts delivered as paper-or e-mail-based messages.
- Information: based on customer-provided or estimated consumption data, consumption estimates augmented with customer-provided information about household appliance stock.
- Difference from Category 2: customer provides information to web server, information provided back is based on estimates for an algorithm.
- Example: Microsoft Hohm, Aclara, Apogee Interactive.

Category 4 – Daily / Weekly Feedback

- Frequency: more frequently than monthly, usually next-day (not real-time).
- Type of Information Provided: Can be basic premise-level kWh/cost information, or additional premise-level comparative information as in Category 2.
- Medium: usually e-mail- or web-based.
- Information: based on metered data (often interval metering data reported on a day-lagged basis to the customer).
- Difference from Category 2: more frequent feedback.
- Difference from Category 3: based on measured data (not calculated estimates), customer does not have to input household information to get results..
- Example: Google PowerMeter (when used with a day-lagged meter data management systems).

Category 5 – Real-time Feedback

- Frequency: real-time or near-real-time (i.e., 15-minute of less (which generally is determined by the meter configuration, not the display device).
- Type of Information Provided: can be basic premise-level kWh/cost information, or more detailed information as in Category 2/4, can include pricing change signaling in a dynamic pricing environment (e.g., change in color when price changes).
- Medium: web-based (or PC- or EMS-based) or through a stand-alone display device.
- Information: based on data transmitted wirelessly from the meter (standard or interval) or via power line carrier using a current transformer connection to circuit box.
- Difference from Categories 2-4: real-time feedback.

- Examples: Blue Line PowerCost Monitor, Aztech In-home Display, Tendril Insight, Comverge Power Portal, Google PowerMeter (when used in conjunction with TED display)

Category 6 – Appliance-level Real-time Feedback

- Frequency: real-time or near-real-time (i.e., less than 30 second lag).
- Type of Information Provided: appliance-level consumption information based on measured appliance consumption; can also include premise-level kWh/cost information as in Categories 2-5 above.
- Medium: web-based (or PC- or EMS-based), through a stand-alone display device, or on display from the appliance itself.
- Information: based on appliance level load measurements system (reported in real-time).
- Difference from Categories 2-4: real-time feedback.
- Difference from Category 5: appliance level consumption information (vs. premise-level).
- Example: Energy Hub, Tendril TREE, PowerHouse Dynamics.

Cross-cutting Variables

Applications of the feedback categories depicted in Figure 1-1 can be configured according to a wide variety of design variables, some of which cut across the feedback categories. Consequently, there can be many different variations within each category. These design variations are depicted in Figure 2-1. It is useful to examine these additional variations, as some may be treatments of interest to test in feedback-related research. The protocols that have been developed to be broad enough to accommodate these sources of variation.

Table 2-1
Cross-cutting Variables

	2 Enhanced Billing	3 Estimated Feedback	4 Daily/Weekly Feedback	5 Real-Time Feedback	6 Appliance Level Real-Time Feedback
Control vs. no control					X
Normative vs. historic comparisons	X	X	X	X	X
Measurement metrics (e.g., $, kWh, CO2, cars-removed or trees-planted equivalents)	X	X	X	X	X
Data presentation (e.g., numbers vs. bar charts vs. pie charts)	X	X	X	X	X
Flat vs. dynamic pricing tariffs	X	X	X	X	X
Stand-alone display vs. PC/web portal vs. TV vs. mobile device				X	X
Instantaneous vs. ‘near-real-time’				X	X
Monthly vs. quarterly vs. annually	X	X			
Delivery medium (mail, e-mail)	X		X		
Opt-out vs. opt-in approach	X	X	X	X	X
Data push vs. pull (e.g., e-mail vs. website alone)	X	X	X		
Goal setting vs. no goal	X	X	X	X	X
Pre-pay vs. regular			X	X	X

Electricity Use Feedback: Past Research Findings and Important Gaps

Conventional electricity billing information (Category 1) provides consumers with limited actionable information about the relationship between how they use electricity and the cost they pay for electric service. Consumers use electricity to operate a wide variety of appliances in their homes, but they receive a monthly statement describing the total cost they incurred for all of the electricity consumed over the period, typically a month or more.

Research conducted over the past several decades suggests that providing feedback on household-specific energy consumption to consumers can cause a change in its timing and/or magnitude. Results reported in the literature indicate that providing feedback of various kinds may cause reductions in energy consumption ranging from -6 to 18%.[7] These studies, which were conducted over three decades, involve a wide range of feedback mechanisms (i.e., Categories 2-5), a wide range of experimental approaches, and a wide range of customer

[7] EPRI, 2009

populations that varied geographically or demographically. More recent research, focusing mainly on in-home displays (IHD, an example of Category 5), suggest overall savings may be in the 0 to 5% range.[8]

The results of prior research, while suggestive of the potential for energy conservation resulting from feedback, are inconclusive on the whole; indeed, they tell us relatively little about the ultimate potential of feedback for changing the timing and magnitude of electricity consumption. As a result, many utilities are unclear about what, if any, additional information they should provide to customers as part of the basic electric service package, or through supplemental programs and offerings.

In reviewing prior studies of the impacts of feedback, EPRI identified five principal gaps in the research literature that hinder progress in the wide-scale implementation of feedback mechanisms:[9]

1. Uncertainties arising from study participation – sample designs and sampling procedures in many, if not most, of the early studies of feedback are either too small to reliably describe energy conservation impacts or represent small and atypical consumer sub-populations from which it is impossible to reliably extrapolate impacts.
2. Impacts of specific delivery mechanisms (i.e., Categories 2-5) are not well understood – there have been few comparative studies of the impacts of different mechanisms on similar study populations or differences in implementation costs. The studies generally do not observe the actual behavior that results in a change in energy use. The studies that have been done usually have sample sizes that prohibit careful analysis of the reactions of customer subpopulations to different delivery mechanisms.
3. Persistence of impacts is not well understood – most feedback studies are carried out over relatively short time frames (some less than one year), or in circumstances that are difficult to generalize to the North American situation. As a result, it is difficult to say with certainty whether the impacts of feedback mechanisms increase, stay the same, or decay over time or abruptly when the feedback mechanism is withdrawn.
4. Uncertainties about the interactions between dynamic pricing and feedback – there have been relatively few studies of the impacts of feedback on consumer response to time-differentiated pricing and the results obtained to date appear to be contradictory.
5. Uncertainties about how different subpopulations respond to feedback mechanisms of one kind or another – most of the research that has been undertaken to date has been designed to estimate a single parameter for the population of interest (e.g., change in kWh usage). While the central tendency of the population is of interest in most research, it is only one property of a statistical distribution that can be modified by providing feedback. It is often the case that the statistical distribution of an impact has considerable range, with some population members responding much more strongly than others. So, while the average response to a

[8] An evaluation of a data from 200 homes in Oregon using the Blue Line PowerCost Monitors showed an effect that was not significantly different from zero. An evaluation of a 30,000-home deployment of the same device in Ontario reported results in the 5.2% range. See respectively: B. Sipe and S. Castor. *The Net Impact of Home Feedback Devices. 2009 IEPEC Proceedings* 2009. and G. Rossini. *Hydro One: In-Home Real Time Display: customer feedback from a 300,000 unit deployment.* 2009. Presented at the Conferences Connect Home Energy Displays Conference, April 2, 2009, Orlando Florida.

[9] EPRI, 2009

given feedback mechanism might be 5%, it is possible that certain subsets of the population respond dramatically more than the average, which would indicate and others respond very little, if at all.

There are presently dozens of products and services designed to provide various kinds of feedback to consumers, and more are on the way. In addition, as of this writing, there were more than 30 pilot studies under way across the U.S. and Canada to test various feedback mechanisms[10]. The questions that must be answered in the very near future about evolving feedback mechanisms are:

1. Do feedback devices and services actually cause electricity consumption to change?
2. Does the degree of change vary across of feedback mechanisms?
3. What other aspects of consumer behavior (e.g., satisfaction with service) are affected?
4. What are the likely participation levels in feedback program under real world operating conditions?
5. Does dynamic pricing complement or compete with the impact of various feedback mechanisms?
6. Do impacts of feedback mechanisms vary across customer segments (e.g., lifestyle categories, income, household family structure, etc.)?

Answers to all of the above questions impact the cost-effectiveness of potential feedback alternatives. Because there are billions of dollars at stake in the decisions to purchase feedback technologies and services, it is necessary that they be answered conclusively. What is meant by the term conclusively? A conclusive research finding is one for which the observed effect of the feedback mechanism (e.g., change in energy consumption) is known to have been solely caused by the feedback mechanism of interest and is not an artifact of the research design, the result of confounding effects, or simply coincidence.

A conclusive research finding is one for which the observed effect of the feedback mechanism (e.g., change in energy consumption) is known to have been solely caused by the feedback mechanism of interest and is not an artifact of the research design, the result of confounding effects, or simply coincidence.

Recently funded stimulus projects will make possible feedback pilots to over a million consumers over the next three years. It is not responsible to settle for results that are merely suggestive. This burst of activity should be directed to specifically and purposefully to clarify how feedback works and to quantify the impacts that result. A set of universally applicable research protocols are a step in that direction.

[10] *Electricity Use Feedback Pilot and Research Activity*. EPRI, Palo Alto, CA. 2009. 1018979.

3
THE ELEMENTS OF EXPERIMENTATION

Conclusively demonstrating that consumer behavior has been changed by feedback and measuring the magnitude of that change are virtually impossible without using modern experimental designs. As will be demonstrated below, this is because it is impossible to conclusively demonstrate that any kind of change in human behavior was caused by any kind of mechanism without using experimental or quasi-experimental research techniques. The basis for this statement, as well as a description of experimental and quasi-experimental designs that are useful in experiments regarding feedback, are described in this section.

What are Experiments?[11]

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

These conditions describe the minimum requirements for *conclusively* demonstrating that feedback causes change in the timing or magnitude of energy use.

An experiment is an actively controlled testing situation designed to fulfill these conditions. In an experiment, the researcher controls the circumstances so that the effect cannot occur before the causal mechanism is present, the objects on which the cause is supposed to operate are observed with (treatment) and without (control) the causal mechanism present, and efforts are made to ensure that other plausible explanations for any changes in the objects of study have been eliminated. Experiments can be described as being more or less conclusive depending on the extent to which these conditions are satisfied.

In the empirical world, 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 conditions to bring about a 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. However, for any of these mechanisms to actually cause a forest fire, a number of other conditions must be present. The forest must be sufficiently dry, the ambient temperature must be high enough, the wind must be blowing from the right direction, etc. In practice, the ways in which all the potential causal

[11] This section draws heavily on a recent report written by one of the authors of this report. See Michael J. Sullivan. *Using Experiments to Foster Innovation and Improve the Effectiveness of Energy Efficiency Programs*. California Institute for Energy and Environment and the California Public Utilities Commission's Energy Division, March 2009.

factors interrelate with one another to actually bring about a given effect are often unknown when one is trying to isolate the effects of one of them.

In an experiment, the researcher controls the circumstances so that the effect cannot occur before the causal mechanism is present, the objects on which the cause is supposed to operate are observed with and without the causal mechanism present, and efforts are made to ensure that other plausible explanations for any changes in the objects of study have been eliminated. Experiments can be described as more or less conclusive depending on the extent to which these conditions are satisfied.

in the saem holds for implementing feedback mechanisms to change human behavior. The feedback mechanism is one of a possibly large number of factors that simultaneously influence energy consumption behavior. Consequently, the impacts of a given feedback mechanism cannot be inferred by observing its effects on a single household or even a small number of households that are provided feedback. Instead, it is necessary to design feedback experiments so that observations are taken for a sufficiently large group of households so that the variation in responses (resulting from the interaction of all the causal factors) within the population of interest can be properly observed.

These are called statistical experiments. One approach is to design experiments so that experimental conditions are repeated a reasonably large number of times (e.g., for many different households) and the impact of the change mechanism (e.g. feedback) between time periods can be described in terms of changes that occur in parameters in the statistical distribution (e.g., mean, proportion, standard deviation, shape, etc.) of effects that are observed over the repeated experiments.

There are good experiments and there are bad experiments. Bad experiments waste valuable time and resources. Good experiments lead us to understand what works and what does not. There is a lot to be gained and lost in the design of experiments. Designs that rely solely on before/after comparisons to measure the impacts of experimental variables have several weaknesses that can render the inferences drawn from them to be unreliable. Below are a number of important things to consider in the design of experiments.

Threats to Internal Validity

During a statistical experiment, a number of things can happen that can result in changes in the outcome variable of interest (e.g., annual kWh consumption) that are not a direct consequence of the supposed causal mechanism (e.g., the feedback mechanism). The change in energy usage may look for all intents and purposes exactly like an effect that might have arisen from the supposed change mechanism. For example, in a simple comparison of annual kWh before and after exposure to a given feedback mechanism (this is 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 feedback mechanism, including the following:

1. History – when a difference in the world at two points in time is observed, 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. Weather is an example of a variable that could cause a change in kWh that might mask the effect of feedback or produce the appearance of an effect of feedback when one did not occur. It is also possible

that news stories, media emphasis on global warming, government programs, and a host of other factors may produce observed differences in outcome variables between two points in time, either masking effects that are attributable to feedback or producing effects that look like the effects of feedback but are not.

Internal validity describes the validity of inferences (or conclusions) that are drawn about the relationship between cause and effect observed in an experiment. Threats to internal validity are aspects of the design of an experiment that can cause experimenters to draw erroneous inferences or conclusions.

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, and not the treatment. An example of a maturation process that could influence the results of a feedback experiment is change in the appliance stock. The appliance stock in households will change over the course of a year-long experiment. As additional new appliances are added to the household appliance stock (e.g. market penetration of flat panel TVs) and older appliances are replaced with newer more or less efficient ones, annual kWh will change even if usage behavior remains the same. Over the whole population of interest, this aging process in the population may produce an increase or decrease in annual energy consumption that could mask an otherwise observable effect of feedback or produce an effect that looks like something that might have resulted from feedback, but did not.

3. Testing **–** when we observe a difference in the world at two points in time, it is possible that the measurement procedures used altered the situation. When humans are involved in experiments, it is sometimes the case that they react to the measurement process in ways that produce the appearance of an experimental effect. This is sometimes referred to as a Hawthorne effect – named for a famous operations research experiment in which worker productivity increased significantly when better lighting was installed not because of the lighting improvement, but because they were being observed. Testing effects are obviously possible in feedback experiments because these experiments involve recruitment of consumers into special testing groups in which new technology will be installed and customers will be asked about their experiences with it (possibly repeatedly). These conditions can lead to behavior changes that appear to correspond with the presentation of feedback, but in fact are due to the observation process.

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 in the precision to which it measures the outcome 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. Calibration problems with instruments used to measure energy use or timing are not likely to seriously influence the results of feedback experiments, although the replacement of old meters with smart meters could result in some average change in energy use due to more accurate measurement. However, this problem can occur with survey instruments administered to treatment and control customer because minor changes in instrument design can produce

apparent (reported) differences between observations taken at different points in time that are solely due to respondents' interpretation of survey semantics

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 than the initial, pre-treatment, measurement. 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. Statistical regression could be a problem in a feedback experiment if consumers were recruited for the experiment from the extremes of the distribution of the dependent variable (i.e., those with very high or very low annual kWh usage). This is because there is random error in the sampling and measurement processes. Because the expectation of the error is zero (under random sampling), there is a high likelihood that subsequent observations will be closer to the mean – just as a result of sampling variation.
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. Mortality does not necessarily mean death. It means that some subset of a sample becomes unavailable for observation for any reason between the first and second measurement periods. In studies of utility customers, this often results when consumers move or change addresses or withdraw from the experiment between the initial and subsequent measurement(s). This causes the measurement of the outcome variable to become censored in the post-test period. When this occurs, the estimated change resulting from the feedback mechanism will be biased. The direction of the bias will depend on the way the censored observations are handled in the analysis. If both pre-test and post-measurements for censored observation are excluded from the analysis, the bias will be in the direction of inflating the magnitude of the effect of the feedback mechanism. If only the missing post-test observations are excluded from the analysis, the bias will be in the direction of suppressing the magnitude of the measured effect of the feedback mechanism. Either way, the situation is highly undesirable.

The above inferential problems all occur because conditions other than the feedback mechanism can cause changes in the outcome variables of interest (e.g., annual energy consumption) when the effect is observed and measured by comparing measurements for a single group at two points in time (before and after exposure to the feedback mechanism).[12]

The above threats to the internal validity of an experiment can be eliminated by changing the design of the experiment so that instead of comparing the reactions of a single group of consumers at two points in time, the impacts of the experimental variable are observed by comparing what happens to two different groups employing random assignment – one exposed to the feedback mechanism and the other not. If the groups are similar, they will experience the same history, mature at the same rate, react to testing and instrumentation in the same manner, etc.[13] In other words, all influences affect both groups equally except for the treatment. In doing

[12] There are circumstances where these problems can be overcome with a single measurement group (i.e., a repeated measures design). Such designs are appropriate when the feedback mechanism is expected to be applied periodically (e.g., Orb used to signal critical peak days). In most instances these designs will not be appropriate for evaluating feedback mechanisms. For a more in depth discussion of these designs, see Sullivan, 2009.

[13] If an experiment is run over several years, it may be necessary to carefully monitor and ensure that control and treatment groups do not change in significantly different ways due to unanticipated differences in customer churn or

so, the threats to experimental validity described above will be eliminated. Of course, this is a very big "if".

The drawback to inferring cause from differences between groups is that the groups may not have been exactly the same to begin with. If they were not, then any observed difference between them could simply reflect the pre-existing difference. This last major threat to internal validity is called selection:

7. <u>Selection</u> – this occurs when groups for which a comparison is being made (experimental vs. control) were different in a systematic way before the measurement was taken. 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. As will become apparent below, because it will often be impossible to randomly assign consumers to treatment and experimental groups in feedback experiments, selection is a potentially very important source of inferential error that must be controlled in feedback experiments.

The above seven problems are what have been described as threats to internal validity. They are plausible ***alternative*** explanations for why a difference might be observed at two points in time (before and after exposure to an experimental condition) for a single group, and for why a difference between two groups exposed to a given experimental condition might occur. Establishing experimental procedures that ensure internal validity is a critical requirement in experimentation. Experiments that are not internally valid (i.e., methodologically flawed) are generally not useful because they do not conclusively show that the experimental variable (feedback) is the sole cause of a change in the outcome variable (e.g., kWh usage). They are, at the minimum, a waste of time and money. They can lead to more damaging outcomes if the results confirm some prior expectation of the result and therefore are readily accepted without additional verification..

Threats to External Validity

The central purpose of most efforts designed to assess the effectiveness and costs of providing feedback is to develop a reliable assessment of how a larger population of consumers exposed to such feedback will react. The fact that occupants in a college dormitory respond to feedback in a certain manner tells us very little about how residential customers in general would respond. This much is obvious. But, the differences between an experimental setting and the broader population need not be so extreme to cause serious errors in inferring that the results of the experiment generalize beyond the experimental setting.

This is the issue of external validity. The external validity of an experiment refers to whether or not the results obtained can be generalized from the circumstances of the experiment (the study groups) to a broader set of circumstances (e.g. the population of residential customer households). 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. <u>Inadequate Representation</u>. If the persons or objects observed in an experiment are significantly different from those of the population for which the generalization is to be

some other factor. If sample sizes are large, this is likely to be of less concern than if sample sizes are relatively small, where random differences in the maturation or mortality of samples could have significant impacts on average use for the two groups.

made, there is reason to suspect that the causal relationship observed in the experiment may not be a validation of the true situation. If the feedback experiment is not conducted with a representative sample of consumers who are expected ultimately to be the recipients of future feedback programs, there is good reason to suspect that the results of the experiment will not generalize to this the entire population.

2. Heterogeneous Settings. Likewise, it is possible that the experimental treatment works differently in different settings. 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 observed in the experimental setting will not hold. This is not likely to be a serious problem with most feedback experiments where the setting in which the feedback occurs is the household. It would be if the subjects relocate to other premises and are kept in the experiment.

The external validity of an experiment refers to whether or not the results obtained in a given experiment can be generalized from the circumstances of the experiment (the study groups) to a broader set of circumstances (e.g. the population of residential customer households).

3. Temporal Stability. If the treatment or outcome measures are changed significantly, there is reason to doubt whether the causal relationship observed during the experiment will hold under wider application of the feedback mechanism.

It is possible to overcome the first and second threats (differences in persons and settings) by employing random sampling from the relevant population of interest (e.g., persons and settings).

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 organizations comprised of a large number of individuals. 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.[14]

Experimental Design – True Experiments

Minimizing the impacts of the above described threats to internal and external validity is the primary objective of experimental design. Sophisticated thinking about the design of

[14] 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 prudent 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.

experiments began about 75 years ago and over the years has produced a vast technical literature that is impossible to summarize in a few pages. This discussion will only scratch its surface.[15]

Completely Randomized Design

In 1935, Sir Ronald Fisher proposed a novel experimental design that eliminated virtually all of the threats to internal validity discussed above. While it is sometimes impossible to employ this design in practical applications, it is useful to understand how it works, because it is the basis for all modern experimental and informs and directs quasi-experimental designs. Moreover, the idealized design serves as a benchmark for assessing the implications for the credibility and extensibility resulting from deviations from that design.

The most elementary experimental design is called a completely randomized design. It is possible to visualize this design as the four-quadrant table shown in Figure 3-1.

	Pre-Test	**Post-Test**
Treatment Group	T_{pre}	T_{post}
Control Group	C_{pre}	C_{post}

Figure 3-1
Block Diagram of Completely Randomized Experimental Design

In this design, observations are randomly assigned to treatment (e.g., households that receive feedback) and control groups (households that don't). Random assignment to treatment and control conditions 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 the experimental effect because the control group, by construction, 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. In other words, all the factors other than the treatment that influence the outcome effectively cancel out. Internal validity is assured by the construction of the design.

Both the treatment and control groups of observations are measured on the outcome variable of interest before and after the experimental factor is introduced. The effect of the experimental variable is measured as the difference between differences.[16] That is:

[15] Those interested in developing a deeper understanding of the problem should consult a very useful and readable summary of the important elements of this literature: Shadish, Cook, and Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* 2002.

$$\textbf{Effect} = (T_{post} - T_{pre}) - (C_{post} - C_{pre}) \qquad \textbf{Equation 3-1}$$

This straightforward 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. Of course, it rests on the assumption that the experimenter has complete control over the selection and composition of the experimental and control groups and over the presentation of the treatment variable.

This design embodies two core ideas used in the design of experiments, First, the use of a control group that is not exposed to the experimental factor. Second. random assignment of individuals form the population to the experimental and treatment conditions. This design is perfectly applicable to all kinds of feedback experiments, and its application will yield incontrovertible findings. However, there are certain practical considerations that often stand in the way of the use of this simple design in feedback experiments. Before discussing these considerations, a few more core ideas from experimental design will be introduced as they provide importance guidance. After that, practical considerations will be discussed.

As explained above, the timing and magnitude of energy use (e.g., possible outcome measures in feedback experiments) arise from the influence of multiple causal factors. While the process of randomization in a feedback experiment will ensure that factors other than the feedback mechanism do not systematically affect the experimental outcome, the operation of other causal factors in the context of the experiment may produce substantial noise or random variation in the outcome variable of interest.

This statistical noise can mute or mask the outcome of an experiment, resulting in a large variance associated with the measured mean outcome. For example, household electricity consumption often varies dramatically within utility service territories by climate, by type of housing, and by appliance holdings. The variation arising from these factors in some instances may be so large that it is impossible to observe the effect of a small but influential feedback mechanism applied to groups. This common problem has led researchers to elaborate on the randomized design to contain the noise.

Blocking factors that may be useful in feedback experiments include: climate zones (if any), lifestyle categories, housing types, family structure types, and usage categories.

Randomized Block Designs

One way to control for experimental noise is by carrying out the randomized design for blocks of customers stratified according to the variable(s) that is suspected of producing the noise. This is called a randomized blocks design. It involves replicating the completely randomized experiment not for the entire population, but for customers within the different blocks or strata.

[16] In practice, it is sometimes impossible to obtain pre-treatment measurements for a variety of reasons. This can happen in a feedback experiment if equipment like interval metering is required to measure pre-treatment energy use. 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 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.

In general, the variance in the regression adjusted estimators for the treatment and control groups decreases in direct proportion to R^2 for the regression equation. Correspondingly, regression adjusted estimates can become very precise when the regression model explains a substantial fraction of the variation in the measurements.

The randomized blocks design is nothing more than a series of four-fold tables like the one in Figure 3-1 – one for each of the blocks or strata. This approach to experimentation is analogous to stratified random sampling in surveying. It can greatly improve the statistical precision of the estimated effect obtained in randomized experiments, assuming there are significant differences in usage or other factors across the blocks.

The benefits of blocking are two-fold. First, blocking can remove potentially large sources of random variation from the measurement of the impact of feedback, thus allowing for a more precise estimate of the unique effect of feedback mechanisms. The changes in the timing and magnitude of energy use that are expected to occur from some feedback mechanisms are small (i.e., perhaps only 0 - 4%). These relatively subtle effects may require fairly large samples to confidently detect without blocking on significant sources of variation. Second, blocking will allow for meaningful quantification and testing for significant differences in the effectiveness of feedback among specific customer segments that may be important for market planning (e.g., households with air conditioning, those living in extreme climates, etc.)

The effectiveness of the randomized blocks design depends critically on having advance knowledge that the experimental affect varies significantly within values of the blocking factor(s). Blocking on a variable for which this is not true will not reduce the noise in the experiment, will generally lead to lower statistical power, and will raise the cost of the experiment.

Covariance Design

An alternative approach to blocking that does not depend as much on prior information about the effectiveness of a 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. That is, consumers are randomly assigned to treatment and control groups. However, in addition to the outcome variable of interest (e.g., annual kWh), measurements are taken on all of the variables that are thought to influence usage (covariates) prior to commencement of the experiment. Examples of covariates that might influence outcome measures in a feedback experiment include:

- Annual household income
- Housing type
- Lifestyle categories
- Dwelling size
- Family size
- Occupancy pattern
- Family structure
- Appliance holdings
- Climate

The variation in the composition of the groups under study with respect to the above uncontrolled, but potentially powerful, causal factors will produce noise in the measurement of the effect of the outcome variable. However, this noise can be greatly diminished by controlling for the correlations between the experimental outcome variable(s) and uncontrolled independent variables analytically through experimental manipulation.

In covariance designs, the values of the uncontrolled independent variables are observed before the experimental treatment has occurred for both treatment and control groups. This makes it possible to estimate a regression function that predicts the mean of the outcome variable of interest (for example, annual kWh) from the level of the uncontrolled independent variables included in the regression equation for both the treatment and control groups.[17] The resulting 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 to establish the effect, regression adjusted means are compared for the two groups.

Figure 3-2 displays an example of a covariance design in which the outcome of an experiment is analyzed by comparing regression adjusted means. The red crosses display the relationship between the outcome variable (measured on the vertical axis) and the covariates for control group members (measured on the horizontal axis). The red regression line displays the predicted average value of the outcome variable given different values of the regression covariates for the control group. The blue circles and regression line describe the relationship between the outcome variable and the covariates for the treatment group members.

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 be small. In studying consumer behavior related to energy use, covariance designs are extremely useful.

[17] 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.

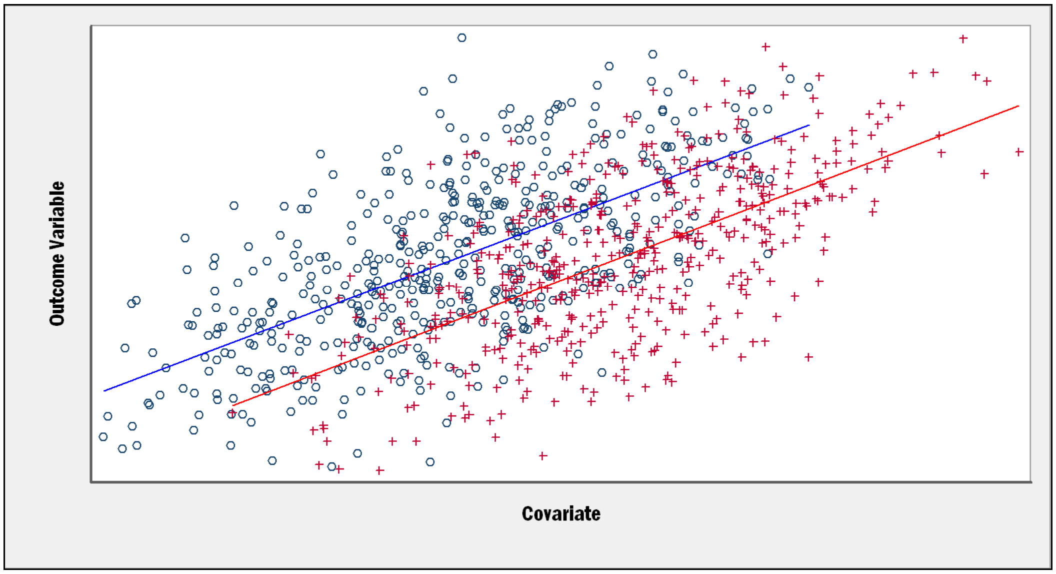


Figure 3-2
Example Analysis of Covariance Adjusted Means

In Figure 3-2, the outcome variable generally increases linearly with the value of the covariate, indicated by the solid blue and red regression lines that are upward sloping and have the same slope. However, there is a difference in the effect of the covariate for the treatment and the control condition represented by the difference where the red and blue regression lines intercept the horizontal axis. Inspecting the graph carefully, it is apparent that while the swarm of outcome value 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 difference isolates the effect of the experimental variable.

The result depicted in Figure 3-2 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 intersect instead of run parallel. This results from circumstances where the effect of the covariates varies with the treatment condition. This is also 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 what results from considering only a simple main effect. The covariance design can be used in virtually any instance where blocking is required, and because it is more powerful, it is recommended over blocking for

It is possible to form a large variety of hybrid experimental designs using combinations of the foregoing core ideas. That is, experiments that involve various combinations of randomized blocks or covariance designs with various kinds of factorial designs can be developed and are frequently used. Readers interested in a further discussion of true experimental designs should consult *Experimental Designs: Second Edition* (Cochran and Cox 1976).

feedback experiments unless the study requires estimation of unique effects within blocks (e.g., low income customer vs. others).

Factorial Experiments

Another useful core idea in experimental design is the concept of the factorial experiment. A factorial experiment is a simple logical extension of the completely randomized design (i.e., randomized treatment and control groups). It is often the case that the experimental factor is not a binary variable (i.e., the treatment was present or it was not). In many cases experimental factors have more than one level (e.g., price differentials in critical peak pricing experiments take on a range of values). In addition, it is sometimes the case that it is desirable to test the simultaneous effects of more than one experimental factor (e.g., variations in the content of the information presented by the feedback mechanism, or where both feedback and education treatments are imposed). In an experiment with more than one factor, it is possible to observe the effects of one of the factors within values of the other factor. [18]

As in the case of blocking, the factorial experiment is a simple extension of the completely randomized design where the columns and the rows in the four-fold table are expanded to accommodate the additional levels within the treatment variable(s).

Factorial experiments are useful because they provide the ability to observe the combined effect of experimental variables on the outcomes of interest. For example, in a feedback experiment it is possible to test the combined effects of dynamic pricing and different information feedback mechanisms. This cannot be done with two side-by-side experiments each separately testing the effects of pricing and information feedback.

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 (e.g., the higher the price the larger the effect of the feedback mechanism). 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 identify optimal feedback system designs, 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 factors in the experiment – independent of the effect of the other. Main effects (e.g., the independent effects of pricing or feedback) are interpretable only if there are no interaction effects.

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 among the variables are interpretable. Interactions involving more than two variables are sometimes difficult to interpret, A factorial design can be used to be sure that the individual treatment effects can be sorted out.

[18] A reviewer pointed out the usefulness of a factorial design to evaluate multiple interventions in a single experiment.

An Important Cautionary Note on the Human Subject Factor

In experiments involving humans, it is often the case that experimental subjects withdraw from the experimental condition to which they have been assigned before the conclusion of the experiment – sometimes even before they have exposure to the experimental treatment. This can occur for all kinds of reasons unrelated to the experimental treatment (e.g., death, change of residence, or replacement of equipment under study). It should be expected to occur in approximately the same proportions for the treatment and control conditions. When this is the case, there is no cause for concern.

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 here. A very comprehensive discussion of these designs is presented in Shadish, Campbell and Cook (2002), and readers wishing to understand the available range of such designs should consult that text as a beginning point.

However, subjects can also withdraw from an experiment in response to the treatment (i.e., they are adverse to it). If this occurs, it can cause serious misinterpretation of study results. If it is suspected that subjects are withdrawing or will withdraw in response to the treatment, it is appropriate to analyze the data from the point of view of the experimenter's intention to treat in addition to the actual exposure to the administered treatment. In an intention to treat analysis, data for all subjects assigned to the treatment and control conditions are included in the analysis (including those who withdrew from the experiment). Since this approach includes parties who were not fully exposed to the treatment, it will generally reduce (properly) the effect of the experimental treatment from what would be observed if only those that stayed in the experiment were analyzed. An excellent discussion of intention to treat analysis and its consequences is found in *What is meant by intention to treat analysis?* Survey of published randomized controlled trials. Sally Hollis and Fiona Campbell (1999).

Experimental Design – Quasi-Experiments

All of the true experimental designs described above have in common the fact that experiment participants (e.g., households) are randomly assigned to treatment and control conditions. For many, if not most, purposefully designed experiments concerning the impacts of feedback, it should be possible to randomly assign observations to experimental conditions using true experimental designs. These designs are definitely preferred over the less robust alternatives discussed below. Every effort should be made to adhere to randomized design principles to ensure that the results are not misleading. Nevertheless, practical considerations will sometimes make the use of true experimental designs impossible and thus it is necessary to discuss practical, second-best, and therefore less desirable, alternatives.

It is not always possible to randomly assign observations to treatment and control conditions. For example, it is impossible to use random assignment when exposure to the treatment condition of interest is compulsory (everyone is required to be exposed to the treatment), or when observations have the ability to select whether or not they are subjected to the experimental condition. These problems commonly occur in experiments conducted in the utility research environment. Examples are pricing experiments where customers volunteer to participate, even if that is accomplished through an opt-out enrollment process.

When random assignment to treatment conditions is impossible, the design of experiments is much more complicated than it is with true experiments. 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 control and treatment effects are compared. 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.

It should be clear that the decision to abandon random assignment can have profound consequences for the internal validity of an experimental design. It places a much heavier burden on the researcher to show that the study's findings are not the result of some unknown and uncontrolled difference between the treatment and synthesized control groups. It can be the first step down a slippery slope that leads to an endless and irresolvable debate about the veracity of the study's findings.

There are several types of quasi-experimental designs that may be particularly important in feedback experiments. They vary according to their robustness (the extent to which they can achieve the credibility of a random experiment) and difficulty in their execution. They are:

- Regression discontinuity designs
- Non-equivalent control groups designs
- Interrupted time series designs

Regression Discontinuity Design

The most robust of the quasi-experimental designs is the regression discontinuity design. In this design, observations are assigned to experimental conditions based on their score on an constructed variable. That variable, called an interval level indicator, represent divides into equal parts the range of possible values (e.g., Fahrenheit temperature, household income, dwelling size, altitude, kW demand, kWh, etc.). As a result, experimental subjects can be assigned a specific interval based on the level of the value that applies to them. In a regression discontinuity design, everyone above or below some point (the discontinuity) on the selected interval scale is assigned to the treatment group, and everyone else is assigned to the control group.

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 systematically with the outcome variable at all. It doesn't matter. In fact, it can be shown that the completely randomized design is just a special case of the regression discontinuity design where the assignment variable is a random number (e.g., everyone above a certain point on the random number distribution is assigned to the treatment group and everyone else to the control group). The impact of the experimental variable in a regression discontinuity design is observed by examining the difference in the regression lines for the assignment variable at the value where assignment was determined.

To see how this works, examine Figure 3-3. It displays two examples of the results of a regression discontinuity 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. The regression line continues uninterrupted or transformed at the assignment value (as indicated by the vertical line in the center of the plot), so there is no discontinuity, which indicates that there is no discernable difference between the treatment and the control groups.

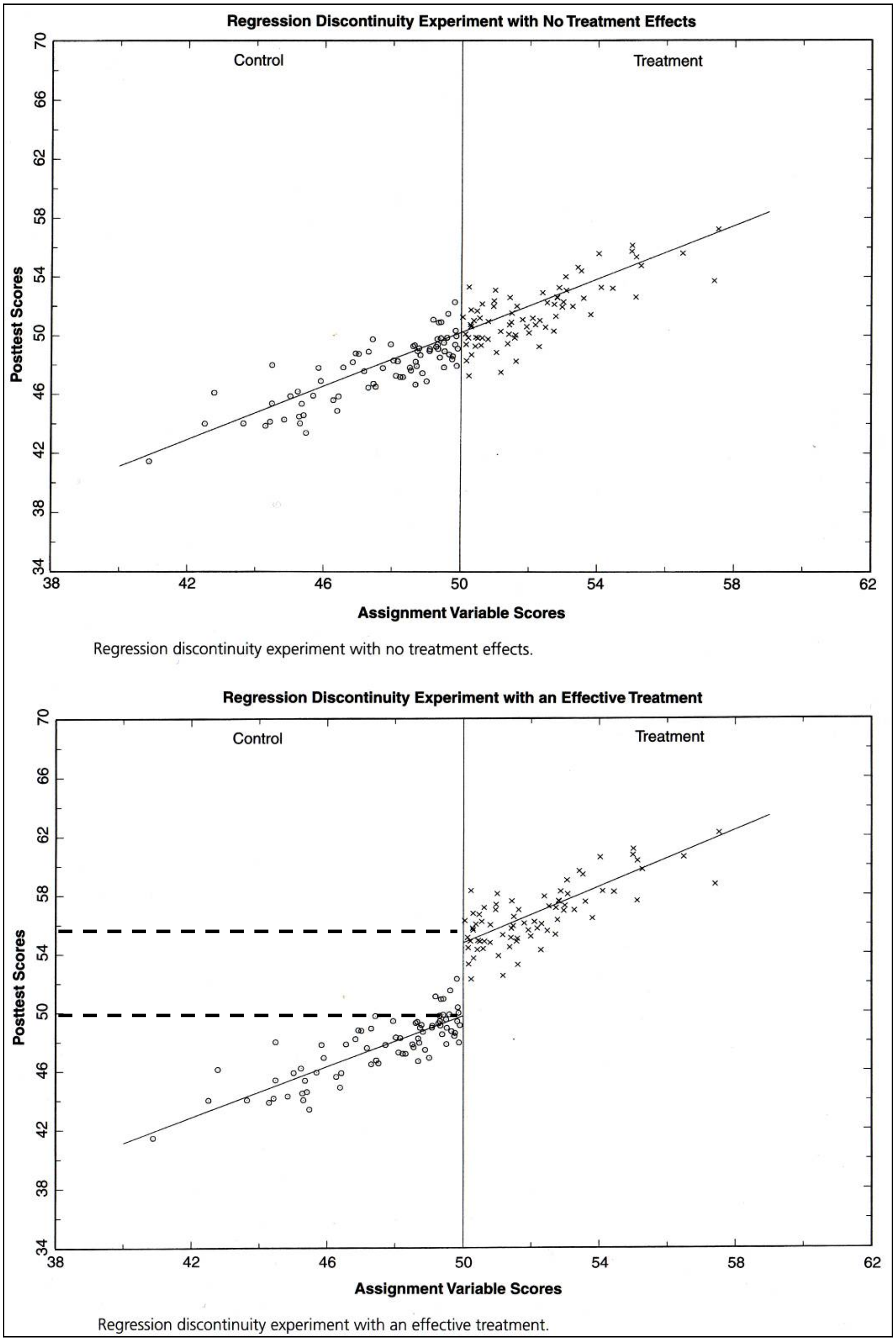
Now compare the regression relationship in the top panel with the one in the bottom panel. Notice the discontinuity at the point on the assignment scale at is again at a value of 50. The difference in the post-test score values at the intersection of the two regression lines depicted in the bottom panel is the effect of the treatment. This effect is illustrated in the figure by the difference on the horizontal axis between the projections of the two intersection points on the vertical discontinuity indicator.

For purposes of reference to an experimental design, in a completely random design the two regression lines would be parallel to the horizontal axis, and the treatment difference would be calculated in the same way.

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 true experimental design in terms of its internal validity. That is, it controls all of the possible alternative explanations for the observed program effect. However, there are certain important caveats that must be met to justify using this design:

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 among experiment subjects 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 estimated regression is specified as linear, but in fact the underlying, predicate relationship is not, the regression discontinuity analysis may incorrectly interpret the point of inflection on the non-linear function as a discontinuity, resulting in a serious estimation error.
4. Likewise, if the treatment interacts with the assignment variable, so that the slope of the regression line changes at the assignment variable due to the treatment effect (causing a jackknife shaped function), and the function is not properly specified as such, this will cause a serious error and one in which the effect of the experimental treatment will be seriously underestimated. Protecting against this possibility requires estimating non-parametric (non-linear) regression functions, which imposes an additional complexity.

Of course, the regression discontinuity design is only achievable where an arbitrary assignment to the experimental conditions is permitted. The assignment should not be related to the treatment, but in fact independent of it. This is not always the case, so other, less robust techniques, may be necessary.


Regression Discontinuity Experiment with No Treatment Effects

Regression discontinuity experiment with no treatment effects.

Regression Discontinuity Experiment with an Effective Treatment

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.

Figure 3-3
Examples of Treatment Effects in a Regression Discontinuity Design

Non-Equivalent Control Groups Design

When random assignment to treatment and control conditions is not possible, another alternative is to try to create a pseudo-control group and analyze the data as though the control group was randomly assigned. This is called the non-equivalent control groups design.

The objective of this approach is to create a non-equivalent control group that is as similar as possible to the treatment group formed by volunteer participants. Non-equivalent control groups are created by selecting control group members from the same population (e.g., neighborhoods, regions, cities, rate classes, willingness to participate in a study, etc.) from which the treatment group came based on their similarity to members in the treatment group. The idea is to sample households from the same population from which the treatment group was selected that are as similar in known respects as possible to the households 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 (or paired matching).

All matching protocols are inferior to randomization in that one can never be certain that the effort to create a matched sample was successful.

Matching is a very old idea and scores of slightly different matching procedures have been tested over the past several decades. Its use is highly controversial for reasons discussed below. The following major types of matching have been used historically:

1. <u>Exact matching</u> – each observation in the treatment group is matched exactly with one member of the control group. In feedback experiments, households could be matched on a number of criteria. For example, each treatment household in a feedback experiment could be matched with a control household randomly selected from the population of non-participants having the same annual energy consumption, or each treatment household could be exactly matched with a control household randomly selected from the neighborhood in which the treatment household is located. The former involves matching based on the variable of interest, the latter matched based on assumed covariates, other factors that should explain energy usage.
2. <u>Caliper matching</u> – each observation in the treatment group is matched within a range inhabited by one member of the constructed control group. This method employs the same basic logic as exact matching except the match is not exact, but is found within a range on the control group. This sort of matching could be applied to feedback experimentation in the same manner as described for exact matching.
3. <u>Bracketed matching</u> – 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 on the matching variable. Again, this is the same basic logic as exact matching except the match is with two observations in the control group. The matching logic would be applied to feedback experimentation in the same manner as described for exact matching.
4. <u>Multivariate index matching</u> – 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, which ideally are covariates. It should be obvious in Option 1 (exact matching) that matching on a single variable does not guarantee that households would be matched on any other explanatory variable, which may introduce

bias at the onset. Multivariate index matching seeks to overcome this problem by matching on an index variable containing the weighted average of potential matching variables, in effect trying to achieve the cancelling out of covariate property of random sampling.

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 feedback experiment) based on information about independent variables (e.g., annual usage, education, household income, etc.) that are correlated with the occurrence of the event in question. Propensity score matching is also used in sequential experimental trials where the experiment is replicated serially two or more times. Propensity scores are created from the first trial to guide the sampling in the second trial to reduce the overall variance and improve the precision of the ultimate estimates of the treatment effect.

Once matching has been completed, the results from the experiment are analyzed in exactly the same manner in which the results from true experimental designs are analyzed. A more robust alternative is to combine linear regression with propensity scores or matching methods.[19] Of course, this approach to matching is only realistic when information is available to calculate the propensity scores for the control group from a compelling set of covariates.

Matching methods by themselves are to be used sparingly because they are prone to the introduction of bias that cannot be anticipated or measured. The calculated estimates of differences (or difference of differences) are biased (they cannot be inferred to reflect the real values) and inconsistent (the variance is large and unknown, so we cannot make statements about the confidence interval around the estimate). These constitute a strong cautionary. However compelling the results based on experience, intuition, or other indicators of a treatment effect, the experiment does not provide confirming and incontrovertible evidence that the observed effect is attributable solely to the treatment.[20]

Interrupted Time Series

Another quasi-experimental design that may be appropriate to feedback experiments is generally referred to as an interrupted time series design. An interrupted time series design 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. 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. In other words, there should be a change in the functional relationship between the treatment and

[19] A reviewer pointed out that this approach has considerable credibility due to its exposition in: Imbens and Wooldridge. 2009.

[20] Some of the reviewers were even more strident in warning experimental designers away from matching to establish controls because there is no way to test to see if the result is indeed biased, leaving that determination to a debate that may be swayed by what some want the results to say. Bias in these cases means neither the level of direction of the effect can be attributed to the calculated effect.

the effect at that time period. In a sense, this is analogous to regression discontinuity, where time is the selection indicator.

This design depends on several important considerations:

1. The onset time of the treatment can be definitively established (i.e., it is definitely known that treatment commenced abruptly 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 time trend influences).
3. If the treatment is expected to have gradually impacted 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)[21].

Interrupted time series designs are subject to several of the threats to internal validity that accompany experimental designs in general. 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 (an exogenous but contemporaneous factor with historical antecedents), or it might have been caused by a coincident 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 has practical applications in analyzing the responses of customers to time varying prices and load management signals. It may also be very useful in analyzing the behavior of customers in response to almost any kind of feedback that affects usage or demand levels. But the intervening conditions described above must either be controlled or not be present. This is not always possible.

An example of how this technique is applied will illustrate its usefulness. Assume that the responses of customers to dynamic pricing signals can be measured at five-minute to one-hour intervals on a daily basis over the course of a season, and that the treatment involves sending price changes to customers routinely over the same time intervals. This is the basis of real-time pricing (RTP) and critical peak pricing (CPP). The effect is measured by observing the impacts of 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.

[21] Autocorrelation is the correlation between levels of 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.

In the parlance of statistics, these designs are referred to as within subjects or repeated measures designs, and they are the state of the art for observing changes in loads and usage in response to price changes. Figure 3-4 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 California Statewide Pricing Pilot.

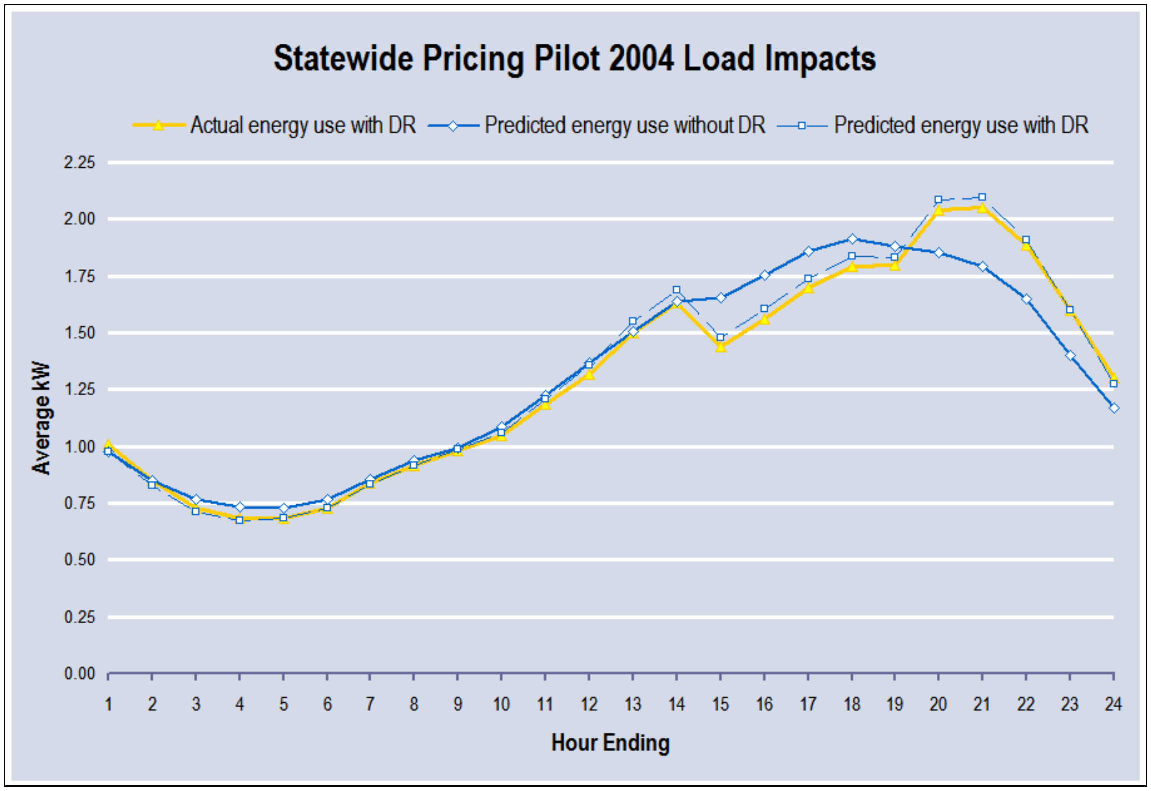


Figure 3-4
Example of Application of Interrupted Time Series Design

In Figure, 3-4 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 in 1) their hourly electric usage on days when the CPP is in effect, and 2) 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.

In conclusion, improving the efficiency of electricity usage is a widely accepted goal in the United States. It can be achieved in a number of ways, including using price incentives, promoting the adoption of more energy efficiency devices, and through the use of on-site renewable generation technologies. All of these are predicated upon demonstrating to consumers that changing their behavior is beneficial to them and has wider implications for the environment and economy. They involve behavioral change.

Feedback may be a potent element of programs to promote the adoption of new technologies and induce consumers to embrace dynamic pricing plans, pay attention to the price changes, and adjust their usage accordingly. But feedback may, by itself, be a principle agent for achieving energy efficiency. It can provide consumers with the information they need to understand when they use electricity, associate a value to that consumption, and adjust their usage accordingly.

Feedback may not only be an important element of any energy efficiency program, it might be the foundation upon which effective and far-reaching programs can be built. If feedback on average reduces household energy usage by 7% to 10%, as some field trials suggest is possible, then it constitutes the largest unified way to achieve energy efficiency goals and aspirations. However, it is almost certainly the case that achieving the likely potential from providing feedback will require significant improvements in our knowledge about how feedback affects human behavior. This is the domain of the social science experiment.

Social science experiments are highly orchestrated inquiries about how humans are affected by changes in their environment. They are designed to remove ambiguity or uncertainty about the relationship between an intervention and the effect that intervention is purported to produce. Usually the intervention or treatment is a remedy or remedial action that is thought to make people better off. In some cases, the purpose of the experiment is to provide information about the efficacy of the treatment as part of making a decision concerning whether the intervention will be offered widely, limited to some subpopulation, or not implemented at all. In some cases, experiments are designed to provide more or less conclusive evidence of causal relationships that theory or prior (uncontrolled) empirical investigations suggest may be operating.

The very fact that an experiment is called for means the stakes are high. It follows then that the experiment's designer should apply scientific principles to ensure that the time and effort involved produce results that are widely accepted for their veracity, even if it fails to conform with what some hoped it would. Only a properly constructed statistical experiment can meet this standard incontrovertibly. Selecting treatments and controls randomly effectively peels away the other influences that can affect the outcome of the experiment and reveals that which can be attributed to the treatment. Attention to design rigor avoids implementation missteps that contaminate the results. A comprehensive analysis plan competently conducted removes any lingering doubts about how to interpret the results in terms of the level of the estimated mean effect, the precision (statistical power) of that effect, and the extent to which inference can be made about the effect of the treatment in wider applications.

There are alternatives that can be substituted for completely randomized designs that, under some circumstances, will yield more or less valid inferences about the effects of experimental treatments. These alternatives are generally more difficult to control and execute than the simpler true experimental designs, and may not eliminate all of the threats to internal validity described in this section. The decision to employ quasi-experimental design has serious consequences. Therefore, it is one that should be taken very carefully and with guidance from experts in experimental design in order to avoid design errors that can lead to erroneous conclusions that become subsequently (and painfully) evident through the adverse outcomes of large programs.

4

RESEARCH DESIGN PROTOCOLS

Content of Research Design

The research design for a prospective information feedback experiment should consist of a number of components, including:

- A technically precise description of the attributes of the feedback mechanism (one of the Figure 1-1 five categories) that will be studied, such as the information content and presentation, delivery mechanism (technology), delivery frequency, etc.
- A concise statement of the research objectives: what you want to know as a result of the experiment along with clear statements of the research question(s) that are to be answered.
- The experimental design that is to be used (e.g., completely randomized design, factorial design, quasi-experimental design, etc.) to achieve interval validity.
- Sampling plan – a plan for selecting persons/households for study from the broader population that may eventually experience the feedback mechanism and for which inference is sought to achieve external validity.
- Recruitment strategy – a strategy for recruiting study subjects in a manner that preserves the validity of the experimental design.
- Length of experiment – the amount of time required for the experimental treatment to take effect and to observe the persistence of and reversibility of its effects.
- Data requirements and data collection methods to be used to observe and record the impacts of the treatment.

> Research design requires balancing budgetary and practical considerations against research design features required to maintain the internal and external validity of the resulting research.

- Delineation of key systems, materials, and support needed to conduct the experiment, including the operational protocols needed to ensure that exposure to the treatment is systematically controlled throughout the experiment.
- Analysis plan – identifying statistical or econometric techniques (e.g., comparison of means, ANOVA, regression, etc.) that may be used to estimate the treatment effects.
- High level budget – an initial budget is developed and compared with available resources, and then subsequent budgets are developed as the research design is modified to align with resource availability.

- High level schedule – a similar process to budgeting, in which the initial schedule typically conflicts with the need for information sooner than is ideal, and several iterations of scheduling and redesign are needed before a final schedule is agreed upon.

To guide research design teams in developing effective plans for carrying out feedback research, this document sets forth ten research design protocols. The ten research design protocols are:

- Protocol 1: Defining Information Feedback Treatments
- Protocol 2: Determining Outcome Variables to be Measured
- Protocol 3: Delineating Customer Sub-segments of Interest
- Protocol 4: Defining the Experimental Design
- Protocol 5: Defining the Sampling Plan
- Protocol 6: Identifying the Recruitment Strategy
- Protocol 7: Identifying the Length of the Experiment
- Protocol 8: Identifying Data Requirements and Collection Methods
- Protocol 9: Meeting Minimum Data Requirements for Cross Utility Comparisons
- Protocol 10: Identifying Key Support Systems and Materials

Research design is a process that involves balancing budgetary and practical considerations against the research design features required to maintain the internal and external validity of the resulting research. Figure 4-1 depicts this process in the context of the protocols recommended for carrying it out. In general, the process starts with the development of precise research objectives. Protocols 1-3 are designed to aid researchers in mapping out the questions that are to be answered by the research in sufficient detail to be able to formulate an appropriate experimental design. Protocol 4 is designed to guide the design team in formulating and describing an appropriate experimental design. Once this is done it is possible to identify the sampling, recruitment, and measurement protocols that will be employed during the experiment. This is the content of Protocols 5-10.

Many key components of the research plan are interdependent – that is, decisions concerning one plan component will affect options and decisions for others. For example, procedures for measuring energy consumption, the timing of energy use, and changes in behavior underlying these factors may entail very different sampling requirements and measurement costs. Likewise, the minimum detection threshold for changes in energy consumption has implications for the sample sizes required in the experimental that can strongly influence measurement cost. The minimum time required for operation of the feedback mechanism in the experiment can influence the complexity of the research design and lead to very different measurement costs.

Knowing that the machines and systems needed to support a given feedback mechanism can be made to work perfectly ignores the elephant in the room – Do they change behavior?

The final research design will be developed through an iterative process in which the objectives and details of the research design are modified to fit within budgetary and other practical constraints. Put another way, it is often the case that the desire for knowledge outstrips the available resources – what you would like to know may involve too many treatments, treatments that are impossible to implement in the institutional context in which the research is being conducted, or insufficient financial resources to support what is agreed upon as the otherwise important research objectives.

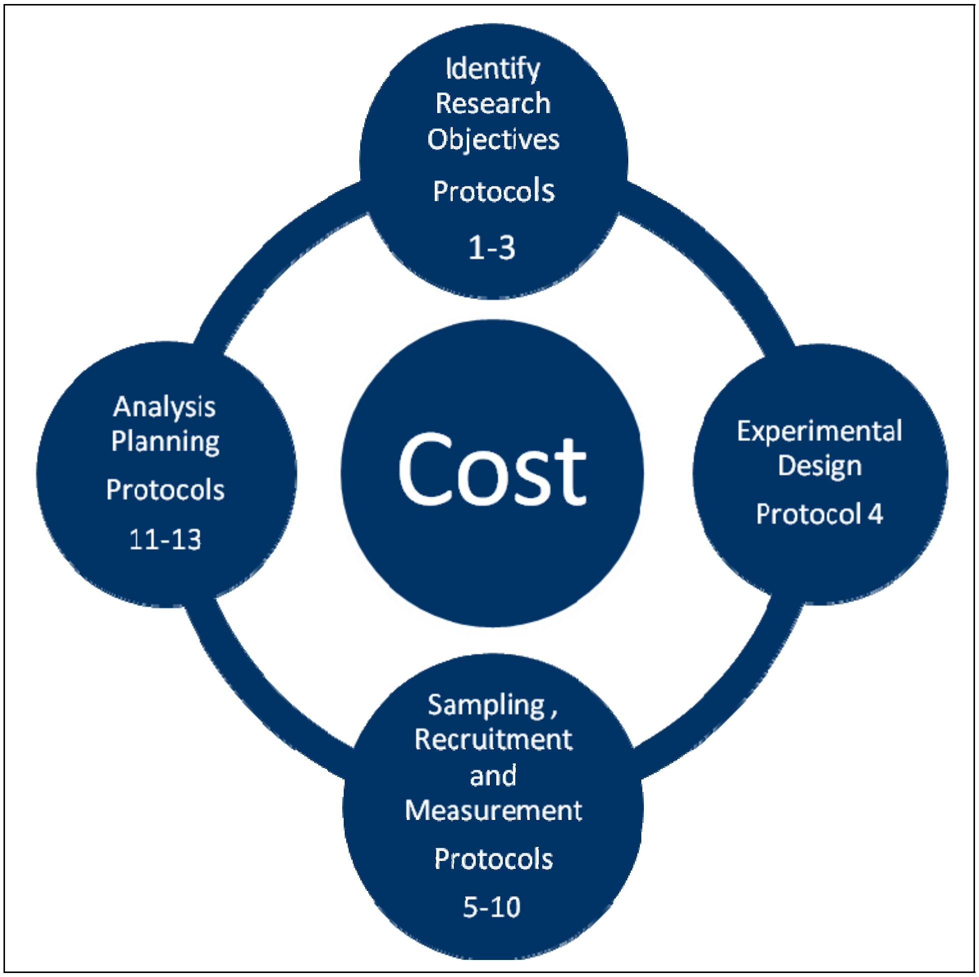


Figure 4-1
The Research Design Process

In such situations, the initial list of "things you want to know" must be paired down to a smaller list of "things you must know" or "things it's feasible to know" in light of practical and/or political limitations. In most cases, the scope of work in the experiment will have to be scaled down to fit within the available resources. As this process takes place it is critical that the research team preserve the design features that are essential for maintaining the integrity of the resulting information. In general it is better to scale back the scope of the research than it is to sacrifice its integrity and jeopardize internal and external validity. For example, if only enough resources are available to test one feedback technology combination conclusively, but two could be tested inconclusively, then the prudent and compelling choice is to test one combination conclusively. Often this requires a strength of conviction about what constitutes a research design, the results of which the designer is willing to back.

The protocols presented here are intended to guide research design concerning the impacts of feedback mechanisms on electricity consumption and electricity usage behavior. As discussed in Section 3, the purpose (aspiration) of experimentation is to conclusively determine whether a given intervention (the treatment) has caused a change in an outcome of interest. Those who are not interested in determining causality need not adhere closely to the tenets and rigor associated with good experimental design. Many pilots that are done by utilities are more focused on determining whether or not a device works, or whether the systems to support it work, than they are focused on clearly measuring the impact on behavior of a new product or service. A lesser objective (it worked or it did not, and not why it worked) lightens the research load, reduces costs, and may be appropriate when impacts have already conclusively been determined.

The research design protocols presented in the remainder of this section operationalize the process of designing an experiment by executing a series of questions designed to help frame decisions for each of the research design components summarized above.

Protocol 1: Defining Information Feedback Treatments

An obvious starting point for research design is deciding what it is that will be tested – that is, defining the experimental treatments. Within the context of information feedback, a treatment can be defined as a bundle of attributes categorized according to the following five important dimensions:

- Information content (e.g., kWh, kW, rate of usage, cumulative usage, rate of expenditure, cumulative expenditure, progress toward a goal point, CO2 emissions from energy use, etc.)
- Information presentation format (e.g., graphic, tabuler, flexible, etc.)
- Delivery channel (e.g., dedicated display device, shared device such as a programmable communicating thermostat (PCT), data pushed to a personal computer, data pushed to some other in-home device such as a television screen, etc.)
- Delivery frequency (e.g., monthly, quarterly, continuous, etc.)
- Interactive features (e.g., “what if” features that allow consumers to determine the impact of a change in behavior on likely energy use or electricity bills, or facilitate goal setting)

A treatment is a specific combination of the above attributes. In essence, it is the mechanism under study. For example, a specific information treatment might consist of a dedicated information display device that reports in near real-time (say every five seconds) the current rate of electricity consumption (kWh/hour), cumulative usage for the billing period (kWh to date), current rate of expenditure ($/hr) and cumulative expenditure ($/month), all presented in tabular format to a separate in-home display devices (IHD) with no interactive functionality.

Of course, a pilot can include non-feedback treatments as well, such as dynamic pricing tariffs, offering energy efficiency program participation, etc. Each of these factors then become treatments in a pilot, allowing for the assessment of individual treatments themselves, as well as the interactions between the feedback and other treatments.

Consumers may respond differently to different content provided in alternative formats through different channels. All of the above design attributes could drive differences in how the information is used and the type of behavioral changes that are driven by the information, and delivery channel could significantly affect customer choice, information usage frequency, and changes in behavior. These types of behaviors are appropriately addressed through the type of experimentation outlined below.

It is possible to construct situations in which a number of different treatments will be simultaneously tested. For example, it might be the case that a single experiment is designed to evaluate the changes in energy use that arise from the use of the first treatment described above, an alternative approach that provides the same data in graphic form, or one that delivers the information to a channel on the customers' television set. Each variation on the content, format, or channel is a different treatment in the experiment and expands the scale of the research effort.

In addition to defining the treatment or treatments that are to be tested, another important initial decision concerns the customer segments that will be included in the test.[22] For example, will the treatments be offered only to residential customers or to non-residential customers as well?

Protocol 1 is made operational through a process described in Table 4-1 that can be used to describe each treatment that will be tested in an experiment and the market segments to which the treatments are administered.

[22] Segmentation here does not refer to any need for segmentation or stratification for the purpose of sampling efficiency. Rather, it is meant to define the customer segments that the treatments will be offered to (e.g., residential, small commercial, etc.). Protocol 3 focuses on sub-segments for which individual impact estimates are desired and Protocol 5 focuses on stratification for sampling efficiency.

Protocol 1

Please complete the following table. When describing the information content that will be made available for each treatment, include a detailed description for Treatment 1 and then define differences in the content between Treatment 1 and the other treatments, rather than repeating the same portions of the description when content overlaps across treatment options. If more than three treatment/segment combinations are to be tested, additional tables should be completed until all treatment/segment combinations are identified.

Table 4-1
Define Treatments

ATTRIBUTE	TREATMENT 1	TREATMENT 2	TREATMENT 3
INFORMATION CONTENT			
Delineate all content for Treatment 1	Detailed description	State the content that is different from Treatment 1	State the content that is different from Treatment 1
INFORMATION FORMAT			
Numerical (toggle through each output)	Y/N?	Y/N?	Y/N?
Tabular	Y/N?	Y/N?	Y/N?
Graphical	Y/N?	Y/N?	Y/N?
Other	Describe	Describe	Describe
DELIVERY CHANNEL			
Dedicated IHD, Professionally Installed	Y/N?	Y/N?	Y/N?
Dedicated IHD, Customer Installed	Y/N?	Y/N?	Y/N?
PCT	Y/N?	Y/N?	Y/N?
Pushed to PC/TV through USB Device	Y/N?	Y/N?	Y/N?
Customer Access through Web Portal	Y/N?	Y/N?	Y/N?
Other	Describe	Describe	Describe
INTERACTIVE FEATURES			
Describe in detail any interactive features provided for each treatment	Detailed description	State the content that is different from Treatment 1	State the content that is different from Treatment 1
DELIVERY FREQUENCY			
Frequency	Describe	Describe	Describe

Protocol 2: Determining Outcome Variables to be Measured

Once the treatments and customer populations of interest have been identified, the next step is to identify the outcomes that will be studied. A feedback experiment can be designed to measure at least the following outcomes:

1. The acceptance rate – the rate at which parties who are offered access to a treatment actually accept it.
2. The impact of the treatment on energy consumption (kW and/or kWh) measured hourly, monthly, or annually.
3. The impact of the treatment on the timing of energy use (diurnal, seasonal).
4. The impact of the treatment on consumer behaviors related to the timing and magnitude of energy consumption – what specific energy using behaviors are changed by how much.
5. The way in which consumers process and use the feedback information (goal setting, understanding impact of changes in energy use by end-user by turning devices on and off, etc.).

The combination of treatments and outcomes of interest have a very significant impact on experimental design and most other aspects of research plan. For example, a relatively simple experimental design can be used to determine the change in average annual electricity consumption that results from exposing customers to a specific treatment. On the other hand, a much more complex experiment is needed to estimate the aggregate impact that would occur for each of several treatments, taking into consideration differences in both average impacts and adoption rates for each option. The more complex the treatments and outcomes under study, the more expensive and difficult the experiment will be to conduct.

Protocol 2 consists of a series of questions that are designed to produce an initial list of outcomes that will be measured in the experiment. As discussed earlier, this list may evolve iteratively if the initial experimental design and the budget required to assess all of the treatments and outcomes of interest exceeds what is available, and therefore not everything of interest can be pursued.

Protocol 2

Please provide answers to the following questions as part of the planning process.

1. Which of the following outcome variables will the experiment be designed to measure? If the outcomes of interest vary by customer segment, indicate the desired outcomes for each customer segment delineated .
   a. Change in annual kWh.
   b. Change in monthly kWh (designate whether for each month or for selected months).
   c. Change in hourly or sub-hourly kWh (designate sub-hourly intervals) for each hour (or sub-hour) for specific, designated time periods (delineate time periods, e.g., all hours in the year, all hours in selected months, all hours on selected days within a month such as system peak days, etc.).
   d. Change in peak demand (kW) for specific, designated times (delineate times, e.g., at time of annual system peak, for each monthly system peak, etc.)

2. Will the experiment seek to identify and quantify the prevalence of the specific types of behavior that change as a result of the treatment? If yes, delineate whether any specific types of behavior are of particular interest (e.g., increase thermostat set point in summer, turn off lights more, etc.).
3. Will the experiment seek to understand how consumers process and use the information being provided to change their behavior?
4. Will the experiment seek to understand the key drivers of customer choice associated with various information options and program/marketing methods? If yes, describe the various marketing strategies/offers that will be tested for each information option and market segment.

Protocol 3: Delineating Customer Sub-Segments of Interest

In addition to defining what the outcome variables of interest are, it's important to delineate whether there are any sub-segments (sometimes called strata) of customers for which separate outcome estimates are needed. For example, is it necessary to establish whether low-income and non-low income customers react differently to the treatment? Is it necessary to ascertain whether impacts vary for households with and without selected end-uses, such as central air conditioning? Is there significant variation in climate across the geographical area under study and is it necessary to take account of this variation in the experiment? These additional requirements impact the experimental design, sample sizes, data requirements, and other key aspects of the research design. Protocol 3 seeks to identify the customer segments for which separate impact estimates are desired, as described in Table 4-2.

Protocol 3

Please complete the following table, indicating the population sub-segments of interest and the a priori assumptions concerning how outcomes for each segment might differ from other segments of interest.

Table 4-2
Delineate Population Segments of Interest

Customer Sub-Segment Description	Hypothesis
Example: Low income consumers	Low income consumers have less discretionary loads and, therefore, are expected to have lower percentage and absolute reductions in annual energy use
(Describe)	(State Hypotheses)
(Describe)	(State Hypotheses)
(Describe)	(State Hypotheses)
(Add additional rows as needed)	

Protocol 4: Defining the Experimental Design

Protocols 1 through 3 are designed to produce a preliminary list of treatments and outcomes of interest and the customer segments that will be the subject of the experiment. Decisions in these areas are necessary, but not sufficient, for determining the preferred approach to conducting the experiment – that is, the experimental design. The ultimate experimental design in almost all cases of social experimentation (in contrast to laboratory studies) reflects a combination of the theoretically correct approach and the practical realities of applied research.

For example, suppose that the objective of an experiment is simply to estimate the short-run change in annual energy use for feedback provided by a specific IHD for a typical residential customer. A completely randomized design involving random selection of control and treatment groups is a simple yet elegant solution to this research problem. Impact estimates can be made by calculating the difference in energy use before and after treatment for both groups and then calculating the difference in the differences for the two groups. This is one of the simplest experimental design cases and it is further simplified by the fact that the only usage data needed, annual kWh, already exists for all customers. As such, there is no need to wait to gather pre-treatment data because every utility already has annual kWh usage on all customers.

Even in this simple case, difficulties can arise in obtaining a truly random sample of treatment customers. A variety of factors can make it difficult or impossible to obtain a random treatment sample. For example, the fact that not everyone will accept participation -- the installation of an IHD. Even if it is offered for free, some customers would refuse it. Or, those who would happily accept an IHD may not be able to be reached for recruitment, are unable to arrange an installation appointment, or accept and after the installation technology problems arise .

If such difficulties themselves are randomly distributed among the population, random selection of control and treatment groups would still be viable. However, if there is some systematic difference between households for whom devices cannot be installed and the general population (e.g., private phone numbers, difficulties in arranging installation for households where both people work during the day, limits imposed by the technology that exclude apartment dwellers, etc.), and these differences affect energy use or the likely change in energy use associated with the device, then a randomly selected control group will no longer provide a suitable comparison group for the treatment population.

There are several solutions to this problem. One solution is to simply treat the parties who were selected out of the intended treatment group as treatment group members for purposes of analysis. Put another way, this approach redefines the treatment group as someone who was offered a device, even if they didn't take it.[23] This is what is called an intention to treat design. This has the advantage of not undermining the basic experimental design.

Another possibility is to over-recruit among parties who agree to participate in the study and then after the recruitment of the total number of participants is completed, randomly assign them to the treatment and control groups or assign them on a first come first served basis controlling for the order variable in a regression. This approach has the advantage of creating perfectly comparable treatment and control groups among the volunteer population.

Finally it is possible to create a matched control group comprised of parties who would have selected themselves into the treatment group, if they had been given the opportunity.

[23] In some ways, this is similar to including treatment customers who received a device but never installed it, if self installed, or ignored it if it was installed by a utility. If the number of customers who agree to participate but for which a device can't be installed is relatively small, and the same percent of customers is likely to be excluded if a program were to be rolled out throughout the service territory, this approach might be both logical and produce more accurate average impacts than one for which it is difficult to pull a matched sample.

There are pros and cons to both approaches that should be weighed carefully when this situation occurs. Basically, treating observations that are not actually exposed to the treatment as though they were, will mute the effect of the treatment (if there is any), and may make it difficult to detect a difference between the treatment and control groups – despite the fact that it provides a statistically accurate measure of the effect of the treatment for the population. This may necessitate a significant increase in required sample size. Creating a non-equivalent control group (comprised of parties who would have selected themselves into the treatment if given the chance) does not suffer this drawback (see Section 3). However, for any non-equivalent control group, there is the possibility of under matching or regression to the mean. The result is fraught with the potential for bias, as was discussed above, arising from the difficulty in ensuring that treatment and control group members are matched on all relevant covariates that influence the outcome variable of interest.

Leaving aside for the moment the question of whether it is generally desirable to use this technique. If the required usage data is available on all households, it would be relatively straightforward from a procedural point of view to construct a non-equivalent control group after observing how the treatment group differs from the general population on important dimensions. However, in some other situations (for example, where pre-treatment data does not already exist) it may be impossible to create a non-equivalent control group based on pre-treatment measurements. This situation is likely to occur when interval meters are required to be installed prior to the treatment period in order to measure the impact of the treatment on peak demand or hourly usage.

Table 4-3 provides some examples showing how the information obtained in Protocols 1 through 3 concerning treatment types, customer segmentation, and desired outcomes drive the development of experimental design. The right hand column in the table also highlights some of the practical challenges that can affect internal validity and possible solutions to these threats.

Table 4-3
Examples of Experimental Design Decisions and Considerations

Treatment	Outcome Variables-Customer Segments	Ideal Design	Practical Considerations (Threats to Internal Validity)
Single Option: Professionally installed IHD	Δ Annual kWh for average residential customer in service territory	Random selection of control and treatment groups	Requires reaching people for recruitment and to arrange appointment for installation. Not everyone will accept device even when offered for free. Above factors will make it difficult to conduct a random treatment sample. Because kWh data are available for all customers, can select suitable comparison group after the fact if it is determined that non-participants systematically differ from participants according to observable factors.
	Δ Annual kWh, maximum kW and hourly usage for average residential customer in service territory	Random selection of control and treatment groups Need to install interval meters and obtain pre-treatment data (assumes advanced metering is not already in place)	Same threats as above. Can't create suitable comparison group after the fact as pre-treatment data would not exist. Could over sample control group to allow for suitable comparison group of sufficient size to be selected after the fact. Could select comparison group after the fact and measure impacts as difference between treatment and control group, not difference of differences.
	Δ Annual kWh for low-income and non-low income customers	Randomized block design with separate treatment and comparison groups for each segment	Same issues and potential solutions as for first example in table.
Multiple Options: IHD, Push to PC and Web Portal	Δ Annual kWh	Factorial design	Same threats to internal validity as first example. May require different comparison groups for each treatment, determined after the fact.

Protocol 4 involves a series of questions that will help guide the construction of the experimental design. It is not possible, or desirable, to dictate the best experimental design for all combinations of treatments, customer segments, and desired outcomes that can arise from

Protocols 1 through 3. Protocol 4 seeks to guide the experimental design process by asking key questions that address not only the theoretically correct design, but also the practical realities that confront real-world social experimentation. When completing these questions, it may be useful to refer to Section 3 of this document as a guide to selecting the experimental design that best supports the treatments, objectives, and practical realities associated with the specific experiment under consideration.

Protocol 4

Please provide answers to the following questions as input to experimental design.

1. Will pre-treatment data be used?
2. Do the appropriate data already exist on all relevant customers, or do meters or other equipment need to be installed in order to gather pre-treatment data?
3. How long of a pre-treatment period of data collection is required?
4. Is a control group (or groups) required for the experiment?[24]
5. Is it possible to randomly assign observations to treatment and control groups?
6. If random assignment is either inappropriate (e.g., if customers are expected to self-select into the program in the future) or impossible to achieve, how will a suitable control group be selected?
7. Using the framework outlined in Section 3, describe treatment(s) and blocks (if any) that will be used during the feedback experiment. This description should be a variation on Figure 3-2 which shows an example of how treatments (and control groups) will be measured for a simple experiment involving two treatments, a control group, and two sampling strata.

Protocol 5: Defining the Sampling Plan

Once the appropriate experimental design has been selected, a sample plan must be developed. Obviously, experimental design and sampling go hand in hand. Sampling is a highly technical problem that should be undertaken by experts. Utilities often staff with some degree of sample design expertise that is able to handle the most straightforward designs. . However, for the more complex designs that may be required in feedback research it may be advisable to acquire the expertise of parties from outside the utility. While an in depth discussion of sample design would lead us far afield of the focus of research design, there are certain critical issues that have to be addressed in any sample design used to study the impacts of feedback on consumers. They are:

1. Are the results of the experiment intended to be extrapolated to a particular population of customers?
2. Are there sub-populations (strata) for which precise measurements are required?
3. What is the absolute minimum level of change in the dependent variable that is meaningful from a planning perspective?

[24] This will almost always be the case, but there are circumstances where other quasi-experimental design techniques can be safely substituted for a control group. See Sullivan, 2009.

4. How much sampling error is permissible?
5. How much statistical confidence is required for planning purposes?
6. Are pre-treatment data available concerning outcome variable(s) of interest?

The answers to the above questions will greatly influence the design of the sample. They cannot and should not be answered by the sampling statistician -- they are program design issues. The answers to these questions must be informed by the policy decisions on which the results of the experiment will rest. That is, they have to be made by the people who will use the information to make decisions given the results. Once these requirements have been developed, a sampling expert can then determine the sample composition and sizes needed to meet the requirements.

Defining the Target Customer Population

If the results of the experiment are to be statistically extrapolated to the utility's entire customer population, then it is necessary to draw a representative (i.e., random) sample from the utility's customer list, and the sample has to be structured so that it is possible to calculate meaningful estimates of the population level impacts using appropriate sampling weights. To calculate weights for purposes of extrapolation, it is necessary to have a list of the members of the population of interest, to sample randomly from that list before assigning customers to treatment and control conditions, and to carefully observe any selection effects that might emerge in the sampling process so that the extrapolation can be adjusted to take account of them.

> If the experimental results are to be extrapolated to the customer population, a representative sample of customers should be subjected to test. Convenience samples should be avoided at all cost.

If precise measurements are needed for specific sub-populations (e.g., low income customers), then it may be necessary to over-sample these customers to ensure that enough observations are present in relevant cells to precisely estimate the impacts of the treatment. These are called sampling strata or blocks as described in Section 3.

Precision of the Estimates

A critical requirement in developing a sample design for a feedback experiment is a clear understanding of the minimum threshold of difference that is considered meaningful from the point of view of those who will be using the results in program planning. As discussed below, the size of the difference that will be considered to be meaningful has profound implications for the required sample size. In general, the smaller the difference that must be detected, the larger the sample size (of treatment and control group customers) needed to detect it. If the cost of a program incorporating the feedback mechanism is known or can be estimated, it is possible to identify the minimum change in energy use that would be required to justify investment in it. For example, suppose a 5% reduction in energy use would be required to justify investment in a given feedback program given estimated program costs in order for the benefits to outweigh the costs. The sample sizes for treatment and control conditions should be set so that a difference of at least 5% can be reliably detected 80-95% of the time.

A related issue that also influences the sizes of samples required in an experiment is the quantity of sampling error that is tolerable from the point of view of planning. This is a slightly different issue from the one discussed above. Samples are sub-sets of the populations of interest and therefore exhibit sample-to-sample variations that cause random error in population parameter estimates derived from them. The magnitude of this sampling error varies with the square root of the sample size. Consequently, the lower the desired sampling error, the larger the required sample size. A sample intended to achieve a +/-1% sampling error requires a much larger sample size than a sample intended to achieve a +/-10% sampling error. This sampling error associated with a given sample is what is often referred to as its statistical precision. It is common for samples to be designed to produce plus or minus 5% statistical precision, but this often cited rule of thumb is not universally applicable to all experimental situations. Ideally, the sampling error that is tolerable should be determined by the economic consequences of the sampling error. Small sampling errors can sometimes cause large variances in downstream calculations (e.g., projected avoided energy production cost or green house gas (GHG) savings). The magnitude of the sampling error that will be acceptable should be determined by examining its consequences for downstream calculations and decisions.

In analyzing the results obtained from a statistical experiment, it is possible to make two kinds of inferential errors arising from the fact that one is observing samples. One can incorrectly conclude that there is a difference between the treatment and control groups when there isn't one (because of sampling variation). This is called a Type I error. Or one can incorrectly conclude that there isn't a difference when in fact there is one. This is called at Type II error. The challenge in designing experimental samples is to minimize both types of errors. This is done by choosing sample sizes that minimize the likelihood of these errors.

Type I – Statistical Significance or Confidence

It is possible to calculate the likelihood of committing a Type I error from information concerning the inherent variation in the population of interest (the variance), the required statistical precision (as described above), and the sample size. This probability – called alpha – is generally described as the level of statistical significance or confidence. It is often set to 5% so that the sample size for the experiment is such that there is no more than 5% chance (one chance in 20) of incorrectly concluding that there is a difference between the treatment and control group of a given magnitude, when there really isn't one. However, as in the case of statistical precision, the selection of alpha is subjective; it depends on the experimenter's taste for risk. It could be set to 1% or 10% or any other level with attendant consequences for confidence in the results. For feedback experiments, it should probably be set to 5%.

Type II – Statistical Power

Type II error is the converse – concluding that the treatment made a difference when in fact it did not. For a given population variance, specified level of statistical precision and sample size, the probability of incorrectly concluding that there isn't a difference when indeed there is a difference is determined by the choice of alpha (the probability of making a Type I error). All other things equal, the lower the probability of making a Type I error, the higher the probability of making a Type II error. In other words, for a given sample size, the more sure we want to be that we are not incorrectly finding a statistically significant difference, the less sure we can be that we have missed a statistically significant difference. The likelihood of making a Type II error can be calculated for a given experiment and generally decreases as sample size increases. The likelihood of avoiding a Type II error is generally referred to as the statistical power of the sample design. The statistical power used in calculating required sample sizes for experiments is subjective and, in modern times, has generally been set at about 90%. That is, it is set so that only one time in ten will the experimenter incorrectly conclude that there isn't a difference of a specified magnitude when indeed there is one. For feedback experiments, statistical power should probably be set at 90%.

The analysis approach used to estimate impacts can also have a significant impact on sample sizes. For example, sampling can be much more statistically efficient if the effect(s) of the treatment(s) are being measured as differences (e.g., pre-test, post-test) of ratios or as regression estimators.

This is true because the variance of these parameters in populations under study is usually quite a bit smaller than the variance of the raw variables, and the smaller the inherent variance of the measurements of interest, the smaller the required sample size. As discussed below, panel regression methods with pre-test, post-test experimental designs can significantly reduce sample sizes.

The following discussion highlights how dramatically sample sizes can vary as a function of customer characteristics (e.g., underlying variation in energy use), research design, and desired levels of statistical precision.

What is to be Measured?

With most feedback experiments, a primary objective is to observe and quantify differences in electricity consumption (and related behaviors) that result from the exposure to feedback provided to consumers. To identify these differences, statistically representative samples of customers can be randomly assigned to different information treatment conditions, and changes in energy consumption can be estimated by observing energy consumption before and after exposure to the information treatments. The sample design question is: how many customers must be assigned to each treatment in order to be able to detect meaningful differences between the groups?

The impacts of feedback are statistically subtle. Results of prior experiments indicate that meaningful differences in electricity consumption arising from feedback variations can range from 1-10%. From the point of view of sample design, these are relatively small differences and, depending on the analysis techniques that are employed, thousands of observations might be required to detect them.

In addition, residential electricity consumption is not symmetrically distributed about the mean or central tendency of the distribution. Instead, the shape of the distribution is right skewed (the right-hand tail extends out farther since the left tail is truncated at zero) causing the standard deviation to be large relative to the mean. The ratio of the standard deviation of a distribution to its mean is called the coefficient of variation. As the coefficient of variation increases, the requirements for sample size increase, all other things being equal.

Table 4-4 illustrates the relationship between detection threshold, coefficient of variation, and sample size requirements for the simplest sample design and analysis approach – independent random samples taken at two points in time and impacts based on a difference of means calculation. The population parameters in the table (i.e., mean, standard deviation, and coefficient of variation) describe actual residential utility customer populations located in different geographical locations.[25]

[25] Calculated using Sampsi formula in Stata.

The first thing to note in the table is that the coefficients of variation vary substantially from location to location. This is due to the fact that some locations are more homogeneous with respect to the distribution of central air conditioning, dwelling size, and other household characteristics. The more homogeneous the population is, the lower the coefficient of variation. It should be evident from the table that all other things being equal, sample size requirements are very sensitive to the coefficient of variation in the population. It should also be evident that any sample design to be used with utility customer populations must take into account the unique coefficient of variation that corresponds to that utility customer population.

A second thing to notice in the table is that depending on the coefficient of variation for the utility customer population for which this study is to be carried out, if independent random sampling were used the sample size required would be between 562,000 and 105,000 per treatment group. This calculation illustrates the folly of employing simple random sampling in studies of utility populations where small differences must be detected. It also shows the powerful impact the coefficient of variation can have on sample size calculations.

Table 4-4
Sample Sizes Required For Independent Random Samples
(Statistical Power 90%, Confidence Level 95%)

Example Sample Sizes Required				
Detection Limit	**Utility 1**	**Utility 2**	**Utility 3**	**Utility 4**
1%	562,727	211,050	397,243	105,024
2%	140,682	52,763	99,314	26,256
3%	62,526	23,450	44,140	11,670
4%	35,171	13,191	24,829	6,564
5%	22,510	8,442	15,891	4,201
6%	15,632	5,863	11,035	2,918
7%	11,485	4,308	8,108	2,144
8%	8,793	3,298	6,208	1,641
9%	6,948	2,606	4,905	1,297
10%	5,628	2,111	3,973	1,051
Mean kWh/mo.	448	565	771	756
S.D.	752	583	1,082	583
Coeff of variation	1.7	1.0	1.4	0.8

The second thing to notice in the table is the variation in sample sizes required to detect different levels of effect. The calculations in the table are illustrative of the sample sizes required to detect differences with independent random samples with 90% statistical power. Notice that the number of observations required to detect a 1% difference at a 90% power level varies from about 562,000 down to about 105,000 as a consequence of the coefficient of variation – a factor of five. It is also important to notice that the sample size requirements vary quite dramatically based on the detection thresholds that are used. Relaxing the desired detection threshold for the size of the effect from 1% to 2%, results in an 80% reduction in the required sample size. So, the decision concerning the required detection threshold has profound consequences for the eventual sample size requirements.

It should be obvious from differences in sample sizes required for different detection limits that the relationship between statistical power and sample size (and cost) is highly non-linear and that careful thought should be given to the required detection limits that are specified in any experiment.

The relationship between required sample size and statistical precision is highly non-linear – the impact of increasing sample size on statistical precision decreases as sample size increases.

Sampling using Panels

Fortunately for most utilities, electricity consumption is measured monthly for purposes of billing and there are extensive historical records describing the monthly electricity consumption for all customers in the majority of populations of interest.[26] This makes it possible to formulate the sample design under the assumption that the data will be analyzed using panel regression techniques, not as independent random samples. These techniques are extremely powerful from a statistical point of view and are capable of detecting small differences between treatment and control groups with relatively small sample sizes (compared to independent random samples).

Panels are independent random samples for which multiple measurements of the dependent variable are available before and after exposure to the treatment conditions. There are two sources of variation in panel measurements – variation arising within the sampled cross-sections (i.e., across the customers due to differences in dwelling size, appliance holdings, and other attributes that are more or less constant over time for each customer) and variation over time within the observations in the panel (resulting from weather variation, experimental effects, and changes in occupant behavior). Because of the need to control for the effects of weather in any given feedback experiment, a minimum of 12 monthly observations of electricity consumption should be observed before the treatment takes place and 12 observations should be taken after the treatment starts (one observation after each time the treatment is given or not given in the case of control groups.) In this way, each observation in the treatment and control groups actually provide 24 observations of household electricity consumption.

Panel regression models derive their power from two aspects of their design. First, there are literally 24 observations for each subject in the experiment, not one – which inflates the actual sample size used in the analysis substantially.[27] Second, the impacts of the experimental treatment are observed within subjects, eliminating variation within the cross-sections. This is particularly useful in observing impacts on electricity consumption because much of the variation in electricity consumption measurements comes from variation across customers in the cross sections.

There are two reasonable approaches to calculating sample sizes required under the assumption that panel regression modeling will be used to analyze the resulting data. First, there are computational formulae available in statistical packages like SAS or Stata for calculating

[26] Exceptions might include new customers with little or no usage history.

[27] Observations within subject are normally auto-correlated and this correlation biases the standard errors that will be obtained from computing formulae that assume independent random sampling. Statistical adjustments for calculating robust standard errors are available and must be used in analyzing panel regression data.

required sample sizes for repeated measures designs.[28] Second, it is possible to use historical population data to generate hypothetical experimental results for samples of different sizes (e.g., 1% differences between treatment and control groups – in a gamma distribution) and then identify the sample size required to detect the simulated difference between treatment and control groups using panel regression models that are expected to be used in the study. Both approaches are reasonable and a discussion of the pros and cons of using them is beyond the scope of this section. For simplicity's sake, conventional computing formulae were used to estimate sample sizes for the experiment illustrated below.

In Stata, the formula for computing sample sizes involving panel regression models requires the following inputs:

1. The desired power of the statistical estimator (the allowed Type II error).
2. The number of pre-test observational periods.
3. The number of post-test observational periods.
4. The desired level of statistical precision (the allowed Type I error).
5. The estimated population mean.
6. The estimated population standard deviation.
7. The expected intra-cluster correlations (for observations in the treatment and control groups).

The basic design of the experiment and requirements for statistical power and precision are predefined and estimates for all of the above parameters can be derived from historical energy consumption information for customers that will be studied. Table 4-5 displays the sample sizes required to detect differences between treatment and control groups using a panel regression model with 12 pre- and post-treatment measurements for each group using Stata's Sampsi algorithm. The sample sizes are for each treatment under study and the utility columns (1-4) illustrate differences in the coefficient of variation (population difference).

The dramatic increase in statistical power using the panel regression modeling technique is evident by comparing the sample sizes contained in Table 4-4 and Table 4-5. Note that, while the whole scale of sampling has been reduced substantially by using panel regression methods with repeated measures, a very significant difference still exists depending on the coefficient of variation in the measurements of the population of interest and the required precision of the estimates.

[28] These computational procedures only take into account the repeated measures aspect of panels and do not incorporate the effects of regression modeling for factors that vary with respect to time that may be modeled with regression. So sample sizes calculated in this manner may be considered to be conservative (on the high side).

Table 4-5
Sample Sizes Required for Panel Regression Model
(Statistical Power 90%, Confidence Level 95%)

Example Sample Sizes Required				
Detection Limit	**Utility 1**	**Utility 2**	**Utility 3**	**Utility 4**
1%	5,061	2,230	3,072	3,595
2%	1,266	558	768	899
3%	563	248	342	400
4%	317	140	192	225
5%	203	90	123	144
6%	141	62	86	100
7%	104	46	63	74
8%	80	35	48	57
9%	63	28	38	45
10%	51	23	31	36
Mean kWh/mo.	448	565	771	756
S.D.	752	583	1,082	583
Coeff of variation	1.7	1.0	1.4	0.8
Intra-cluster correlation	0.9	0.9	1.0	0.8

Given the dramatic improvements in statistical power that can be obtained using repeated measures designs, it seems unnecessary and ill-advised to apply sample design algorithms designed to identify sample sizes required for independent random samples to the design of samples to be used in feedback experiments.

The application of Protocol 5 involves answering a series of questions, as follows:

Protocol 5

Please answer the following questions pertaining to sample planning.

1. Are the measurements from the experiment to be extrapolated to the broader utility population?
   a. If yes, indicate whether the sample will be stratified and what variables will be used in the stratification.
   b. If no, describe the list of customers from which the sampling will be obtained.
2. Are precise measurements required for sub-populations of interest?
   a. If yes, describe the sub-populations for which precise measurements are desired.
3. What is the minimum threshold of difference that must be detected by the experiment?
4. What is the acceptable amount of sampling error or statistical precision and acceptable level of statistical confidence (i.e., 90%, 95%, 99%)?

5. Will customers be randomly assigned to treatment and control conditions or varying levels of factors under study?
   a. If yes, do you expect customers to select themselves into the treatment condition?
   b. If so, how will you correct for this selection process in the analysis and sample weighting?
6. If customers will not be randomly assigned to treatment and control conditions or varying levels of factors under study:
   a. Describe the process that will be used to select customers for the treatment group(s).
   b. Describe the process that will be used to select customers for the control group, and explain why this is the best available alternative for creating a non-equivalent control group.
   c. If no control group is used, explain how the change in the outcome variables of interest will be calculated.

> If they are known, recruitment strategies used in the experiment should mimic those that are expected to be used in the planned feedback program.
>
> If they are not known, variations in recruiting strategies should be incorporated into the experimental design as treatments.

Protocol 6: Identifying the Recruitment Strategy

The primary focus of recruitment planning is to implement a strategy that preserves the internal and external validity of the experimental design. Recruitment strategy is influenced by a variety of factors.

What Strategy to Test

The first consideration is whether or not the recruitment strategy that would be used in full-scale implementation (e.g., direct mail, door-to-door, etc.) has already been determined. If so, then the recruitment strategy for the experiment should mimic this strategy as closely as possible. This reduces concerns that different recruitment strategies might attract different types of participants than would be produced by the recruitment strategy that is ultimately intended to be used.

It is also possible that one of the primary purposes of the experiment is to evaluate recruitment strategy alternatives and identify the most cost-effective approach for purposes of program design, taking into consideration both the number of enrollees as well as the average savings per customer. In this case, the experimental recruitment strategy would use various approaches and carefully monitor the differential take rates and differences in customer characteristics and impacts associated with each approach.

To Stratify or Not?

Another consideration is whether the sampling plan that is to be used involves stratification and, if so, whether the stratum characteristics are known a priori or must be determined as part of the recruitment process. For example, if stratification is based on annual usage, that information is known for all customers a priori and samples can be drawn and progress toward sample goals easily tracked based on usage data. On the other hand, if stratification is based on income or the presence or absence of certain appliances, for example, information on these variables must be gathered during the recruitment process. The need to pre-screen observations on stratification variables that may not be known prior to contacting the customer may require more flexible and interactive recruiting processes that involve direct customer interaction (i.e., telemarketing or door-to-door), even though the eventual marketing effort may be through direct mail.

Who's Eligible?

Another consideration is whether there are eligibility criteria that must be met in order to participate in the program of interest. For example, if one of the purposes of the experiment is to determine consumer preferences for dedicated IHDs vs. devices that push real-time data to a personal computer (PC), consumers who do not have a PC are not eligible for the second treatment. In order to determine preferences for the two options among the same population, the recruitment process must screen for PC ownership.

The cost per sample point will influence the design of the recruitment process. If the cost per sample point is high (e.g., because meters and/or IHDs must be installed for each participant), the cost of exceeding the required sample size in each stratum will be relatively high and the recruitment strategy must carefully monitor progress toward sample targets and include a means of stopping enrollment once the target within any given cell is reached. On the other hand, if the cost for each sample point is low, the cost of turning people away (e.g., customer ill will, monitoring costs, etc.) might exceed the cost of accepting everyone into the study.

How Long Should the Recruitment Effort Run?

The time available to complete the study is another consideration that will affect the design of the recruitment process. Certain recruitment processes take longer than others and the ideal process may not fit within the time frame available to complete an experiment. For example, direct mail recruitment into a program or experiment can take several months of calendar time in order to achieve maximum enrollment, as multiple waves of marketing materials must be sent to targeted households.

The same level of enrollment might be achievable through telemarketing in a few days time. If the time available to field an experiment is short (e.g., because a decision to go forward was delayed and there is a need to get into the field prior to the summer season), it may not be possible to recruit customers through direct mail even if the ideal experimental design calls for it. Under these circumstances, if possible, it would be important to factor into the recruitment process and overall experimental design a calibration step that would allow one to determine if customers recruited through direct mail and telemarketing differ in important ways that might affect the impact estimates from the experiment. For example, even if direct mail would not allow for sufficient recruitment to occur prior to a seasonal deadline, one could still recruit a small group of customers through direct mail and compare the observable characteristics of these customers with the group that was recruited through telemarketing to determine whether there are differences in observable characteristics that might affect the impact estimates.

Are Incentives to Participate Appropriate?

Careful thought should be given to the decision as to whether incentives should be used in experiments regarding the impacts of feedback on the timing or magnitude of electricity consumption. Economic incentives can be offered to encourage participation in experiments. They can be used to enhance the likelihood that consumers continue to participate in programs and they can be offered to enhance the magnitude of consumers' responses to feedback. The use of all of these kinds of incentives in feedback experiments is appropriate only if they are being considered (in fact tested for efficacy) for use in the ultimate implementation of utility program designs under study.

Indeed, the identification of the magnitude of the impacts of incentives on each of the above kinds of behavior are perfectly legitimate areas of investigation in the event that they are being considered as part of an eventual strategy to enhance the effect of feedback. However, because incentives can have powerful effects on experimental outcomes, they should not be used if they are not expected to be part of any future program. Doing so may seriously undermine the external validity of the experiment and lead to potentially erroneous conclusions about the likely future participation in and impacts of opt-in feedback programs.

> If recruitment methods other than those that will be used in the full scale rollout of the program are used in the experiment, it is necessary to calibrate the response rate obtained in the experiment to the response rate that would be obtained using the recruitment method that is expected to be used in the full scale program.

Because many of the experimental designs that will be employed in feedback experiments will require acquiescence of the subjects to the experimental exposure, careful attention should be paid during the recruiting stage to collecting information that will be needed to describe the impacts of selection and to carry out intention to treat analyses when the experiment is concluded. This means that the characteristics of those who reject the treatment and measurements of the outcome variables of interest must be collected for these customers. Collection of this information for members of the treatment group should always be done as part of the experimental design.

Finally, in experiments that are designed to saturate large segments of a given geographical market there is the risk that messages transmitted to the treatment group will inadvertently be received by the control group. This can occur when treatment and control group members are intermixed in neighborhoods. The complications associated with geographically isolating treatment and control groups probably introduce more risk of experimental failure than communications among the two groups. However, given the possibility of contamination between treatment and control groups, surveys of control and treatment group members should include questions designed to detect and control for this possible contamination.

Protocol 6 contains a series of questions designed to elicit information useful for planning the strategy that will be used to recruit customers in a manner that supports the integrity of an information feedback experiment.

Protocol 6

Please answer the following questions pertaining to recruitment.

1. Is the approach to recruitment for a full-scale program that might ultimately be implemented known with certainty?
   a. If yes, does the project timeline allow for experimental recruitment to be done in the same manner as the planned recruitment?
   b. If yes to Question 1a, what is the recruitment approach that will be used (e.g., direct mail, telemarketing, door-to-door, etc.)?
   c. If no to Question 1a, what recruitment options fit within the available timeline?
      i. What are the potential differences between customers who would be expected to enroll through the long-run recruitment process and customers who would likely enroll through the process that will be used in the experiment?
      ii. Is it possible to recruit a calibration group using the long-run recruitment approach even if they cannot be enrolled in time to be used in the estimation sample for the load impact analysis?[29]
2. Is one of the purposes of the experiment to determine what recruitment process works best and, if so, which options will be studied?
3. Does the sampling plan involve stratification?
   a. If so, do data exist that allow for stratification prior to recruitment or does the recruitment process need to gather data on customer characteristics and track enrollment according to these criteria?
4. What eligibility criteria, if any, apply to each treatment option?
   a. For each treatment option that has eligibility restrictions, do data already exist that allow for precise targeting of eligible customers?
   b. If the answer to Question 4a is no, does the planned recruitment approach allow for eligibility screening to occur and be tracked as part of the recruitment process?
5. Taking into consideration the cost of each sample point and any other relevant criteria, how important is it to cut off enrollment as close as possible to the target sample size?
6. If incentives are to be used to enhance subscription, improve persistence, or increase the magnitude of the response to the feedback mechanism, describe the incentives that will be offered and the variations in magnitude of the incentive that will be tested during the experiment.

[29] For example, it might be necessary to recruit by telephone in order to meet a deadline to install meters prior to a summer season when treatments must go into effect. However, in parallel with this effort, it could be useful to recruit a small sample of customers using direct mail, even though it would not be possible to enroll them and install meters prior to the start of the treatment period. The characteristics of this small calibration group could then be compared with those of the group recruited through telemarketing to determine whether there are observable differences in the two groups that might affect the impact estimates obtained from the telemarketing recruitment process.

Protocol 7: Identifying the Length of the Experiment

The period of time over which an experiment will be run almost always involves a compromise between what is optimal from a research perspective, what is required in terms of regulatory and/or management needs, and available budget. As a general rule, the longer an experiment is allowed to run, the more likely it will be that impact and enrollment estimates obtained from it will reflect the long-run potential of the treatments being examined.

Even if an experiment is primarily focused on understanding changes in usage behavior, it takes consumers time to process the information being provided and perhaps to learn new and innovative ways of responding to it. It may require only a few months to observe whether exposure to the feedback mechanism has changed the short term timing or magnitude of energy use. However, this is almost certainly not enough time to determine whether the observed changes persist, increase, or decrease with the passage of time, nor is it long enough to assess potential long-run effects of the feedback, such as the purchase of more efficient appliances.

This is not a minor consideration if a utility is cost-justifying its investment on the basis of the assumption that the changes that are observed during the short experimental period persist throughout the life of the investment. While it is undoubtedly impractical to run an experiment long enough to confidently predict the ultimate persistence of most feedback mechanisms, it is nevertheless advisable to run experiments long enough to allow the trend in persistence to be observed.

Considering the time required for consumers to fashion and implement new household energy usage practices, experiments involving feedback should run for at least two years after the treatment is started. It may be necessary to observe the behavior of consumers for an even longer time period, or to greatly expand the sample sizes for the treatment and control groups (so that more appliance replacements are available for study), if the treatment is intended to cause a change in appliance acquisition behavior – given the relatively slow turnover rate in household appliances.

Feedback persistence is an extremely important policy question. Because behavior is malleable, there is inherent uncertainty as to whether behaviors that are initially adopted are sustained over time, and become habits. The longer the experiment, the more likely it is that any decay in the response will be observed. So, longer experiments are definitely to be preferred to shorter ones.

However, in many cases it will be simply impractical to delay the decision as to whether to implement an intervention until an experiment has been allowed to operate for the duration of the expected life of feedback equipment. In these cases, it will be necessary to forecast persistence based on the trends observed in the experiment. Ultimately, the forecast of persistence will depend on one’s confidence that the trends observed during the experimental trial result from changes in household behavior that are likely to persist over time. One way to make this determination is to put into place a means for determining what actions customers undertook that resulted in lower electricity use, including those that can be attributed to responding to feedback and those that may reflect a change of circumstances external to the experiment (e.g., children moving out or back in). Surveys and other methods of data collection (e.g., blogs, diaries) are required to identify and track such changes. Protocol 7 contains a series of questions designed to determine what would be ideal in terms of the length of time over which an experiment should be run, as well as what is necessary to meet practical considerations. As a practical matter, cost and the need for answers sooner rather than later very often constrain the time period to something less than ideal.

Protocol 7

Please answer the following questions pertaining to the experimental time frame.

1. Is it possible to run the experiment for at least two years?
   a. If no, how will the persistence of the effect be determined?
2. What is the maximum amount of time consumers can be exposed to the feedback mechanism?
3. Do pre-treatment data for the relevant variables already exist or must time be allowed to obtain pre-treatment data?
   a. If pre-treatment data do not already exist, how long must the pre-treatment period be to support the experimental objectives?
   b. If pre-treatment data do not already exist, can the experiment be conducted using only post-treatment data, and what adjustments to sample design will be required to employ a post-test-only design?[30]
4. What is the expected amount of time required for consumers to receive and understand the information being provided to them?[31]
5. What is the expected amount of time needed by consumers to implement behavioral changes in response to the information provided?
6. How long between the time when a consumer implements a change in behavior and when the feedback associated with that change is likely to be delivered to consumers?[32]
7. What is the minimum amount of time the effect of the feedback mechanism must persist to cost-justify investment on the part of the utility?
   a. If the duration of the experiment is shorter than the expected useful life of the measure, how will the determination be made as to whether the effect of the feedback persists long enough to be cost effective?
8. Is the feedback mechanism expected to affect consumers' decisions about the energy efficiency or demand responsiveness of new/replacement appliances?
   a. If yes, how will the impact of the feedback mechanism on this behavior be measured?
9. How much time is needed between when the research plan is completed and approved, and when treatments are in place for experimental participants?
10. How much time is required between when the final data are obtained from the experimental observations and when the analysis can be completed?

[30] Put another way, are pre-treatment data essential or is there a work around that can be used if the experimental time frame does not allow for the collection of pre-treatment data?

[31] For real-time feedback, this time period is likely to be measured in days. For monthly information provision, it could take several months before consumers would receive sufficient information feedback to factor it into their usage decisions.

[32] With real-time feedback, the time required for consumers to observe the impact of a change in behavior is almost instantaneous whereas for monthly feedback, it may take several months to see the affect of a change.

11. What are the drop-dead dates for when draft and final results from the experiment are needed?

Protocol 8: Identifying Data Requirements and Collection Methods

A critical part of research design is delineating all of the data that will be required to determine whether changes occurred in the measurement variables identified in Protocol 2 for each segment and sub-segment of interest and determining the best way to collect the necessary data. It is also essential to identify the timing associated which each data collection step.

The data requirements evolve directly from the experimental objectives and target markets outlined in Protocols 2 and 3. If the only focus is on the energy or demand impacts associated with an information feedback treatment, there may be few data requirements beyond simply gathering the relevant data from a utility's customer information system (CIS). On the other hand, if information is sought on the behavioral changes that customers make and/or on how they use the information provided through each treatment to change their behavior, the data requirements and collection methods may be extensive.

As with most other elements of research design, developing a data collection strategy typically involves tradeoffs between what is desired, the level of intrusiveness associated with obtaining certain types of information, the accuracy of various methods, and as always, cost. For example, a relatively low cost option for learning about behavior change is to ask treatment customers through a mail or telephone survey, at the end of the treatment period, what types of behavioral changes they made as a result of the information feedback received. However, self reported information is less accurate than direct observation of behavior or metering end-use appliances before and after treatments go into effect. The more revealing alternative, end-use metering, is currently quite expensive and somewhat intrusive, and direct observation is very intrusive and probably even more expensive than end-use metering.

An approach that is somewhere in between the extremes of end-of-treatment surveys and direct observation or end-use metering would be to conduct several brief surveys before and during the treatment period. Such surveys may be more accurate than a single survey at the end of the treatment period, as they rely less on recall after the fact than on reporting behavior at the time of the survey. However, the survey burden and cost associated with multiple surveys are higher than with a single survey conducted at the end of the treatment period, and frequent surveys of the same individuals can increase the risk of creating Hawthorne effects.

Protocol 8 contains a table that can be used to delineate the data requirements and collection methods needed to support the research. For each category of information, the table summarizes the specific content of the information needed, the population of interest, the frequency with which data will be collected, the primary method of data collection of source of such information, and any issues that might arise and solutions to address them. Examples of what might appear in the table entries are contained in Sections 8 through 10.

Protocol 8

Please complete the following table delineating the data requirements and collection methods that pertain to the proposed research.

Table 4-6
Data Requirements and Collection Methods

Energy Use	
-Description	
-Population	
-Frequency	
-Method/Source	
-Issues and solutions	
Socio-demographic and appliance data	
-Description	
-Population	
-Frequency	
-Method/Source	
-Issues and solutions	
Energy using behavior	
-Description	
-Population	
-Frequency	
-Method/Source	
-Issues and solutions	
Use of information	
-Description	
-Population	
-Frequency	
-Method/Source	
-Issues and solutions	

Table 4-6 (continued)
Data Requirements and Collection Methods

Weather data	
-Description	
-Population	
-Frequency	
-Method/Source	
-Issues and solutions	
Other	
-Description	
-Population	
-Frequency	
-Method/Source	
-Issues and solutions	

Protocol 9: Meeting Minimum Data Requirements for Cross-Utility Comparisons and Pooling

As discussed in Section 1 of this report, one of the objectives of these protocols is to enable the utility to compare results across experiments and to support collective research by enabling the pooling of data across experiments and utilities. If each utility conducting an information feedback experiment collected a standard set of data coded according to a common format so that individual customer data could be pooled across utilities, it will be possible to learn much more about the effects of feedback more quickly than if results from experiments can not be usefully aggregated.

There are a host of reasons why individual utilities may resist the assembly of useful information about the outcomes of feedback experiments. Data collection can be costly and gathering even a minimum set of data that would support cross-utility comparisons or data pooling may increase costs for some utilities that would not need such data for their own internal purposes. Considerations other than cost are also relevant. Even when two utilities are planning to obtain survey data on the appliance holdings and socio-demographic characteristics of study subjects, there may be reasons why a utility might not want to use a common survey instrument, which renders cross-utility comparisons or pooling a fruitless exercise. For example, it may be more important to each utility to use a survey instrument that was used in prior internal surveys in order to compare the characteristics of study subjects with the general utility population than to use an instrument that enables cross-utility comparisons.

Nevertheless, EPRI believes that the development of standardized data concerning the outcome of feedback experiments should be a priority among its members and that there is value in describing a minimum set of data requirements that would support cross-utility comparisons of results and/or data pooling. Protocol 9 delineates these minimum requirements that could be used by utilities that are interested in supporting more meaningful cross-utility comparisons and/or participating in joint analyses with other utilities.[33]

Protocol 9

In order to enhance cross-utility comparisons of experimental results or to allow for data pooling across experiments, the following data should be obtained for each experimental subject.

1. Designator indicating the treatment to which the observation was assigned (e.g., Treatment 1, Treatment 2, Control, etc.)
2. For customers in all experiments that do not involve interval metering:
   a. kWh usage for all pre-treatment and treatment billing periods for each participant
   b. Meter read date for each billing period
   c. Monthly electricity bill
   d. Tariff designation
   e. Date that treatment went into effect for each treatment customer
   f. Date customer left experiment for each customer that left before the end of the treatment period
3. For customers in all experiments involving demand-metered customers, in addition to all of the data in Question 1 above:
   a. Monthly peak demand
4. For customers in all experiments in which all customers have interval meters:
   a. kWh usage for each hour for the pre-treatment and treatment time periods
   b. Items 1b, 1c, 1d, and 1e

[33] There is a parallel protocol for documentation in Section 7 of this report that sets out minimum reporting requirements that will enhance cross-utility comparisons of output results. These reporting requirements rely on the data delineated in Protocol 9.

5. For customers in all experiments, data on the following customer characteristics:

Variable	Specification
Zip code	5 digit
Date customer entered the experiment	mm/dd/yy
Date customer departed the experiment	mm/dd//yy
Reason customer withdrew from experiment	Text (e.g., deceased, moved, etc.)
Air conditioning systems	Number of central AC units Number of room AC units
Space heating systems	Presence of electric baseboards (Y/N) Number of central heating systems (gas) Number of central heating systems (electric)
Type of space heating system control	Manual Standard thermostat Programmable thermostat
Water heating systems	Electric Gas Solar
Household appliance inventory	Number of the following appliances: Home computers Electric spas Printers Pool pumps Dishwashers Domestic water pumps Clothes washers CRT TVs Electric dryers Plasma TVs Electric cook tops LED TVs Electric ovens
Dwelling type	Single family detached Single family attached (e.g., duplex or town house) Multifamily (e.g., apartment or condo) Manufactured home (e.g. mobile home) Other
Dwelling size	Sq. ft of enclosed area
Number of persons in household by age group	Age 1-6 7-19 20-24 25-60 61-70 > 70
Annual household income	For the year preceding the start of the experiment

Protocol 10: Identifying Key Support Systems and Materials

Another key step in experimental research planning is identifying the systems and materials that will be needed to implement the experiment. Required systems and materials can be expensive and if not properly anticipated, can put an entire experiment at risk of delay or even failure. Making modifications to existing utility systems to support an experiment may be more costly and risky than working around such systems through outsourcing. Even when there is every reason to think that internal systems will be able to support an experiment, there may be value in developing a backup plan that can be used in case internal development falls behind and puts the experimental schedule in jeopardy.

The idea of working around evolving systems may even be useful for metering. Many utilities may be interested in conducting information feedback experiments while new smart metering systems are being deployed, but before deployment is completed. Smart Meter deployment is never random. If a representative sample of a utility's entire customer population is needed for an experiment, it may be necessary to install meters for some customers well ahead of when they would otherwise be scheduled to receive one.[34]

Feedback technology selection is also an important step in research planning. Here too, it may not be necessary or possible to use the same technology in an experiment that might ultimately be deployed to provide information feedback as part of a full-scale program. What is required is that the technology used in the experiment provides very similar content and functionality as the technology that might ultimately be deployed. For example, the focus of an experiment might be on determining whether to provide, at some point down the line, an IHD device that would communicate directly with a new smart metering system using an open communication protocol such as ZigBee. However, ZigBee is still an evolving protocol and may not be finalized for several years. But the means by which real-time information is communicated from a meter to an IHD device is largely irrelevant to accurate measurement of customer impacts. Indeed, IHD devices that require only conventional metering devices are available and may be sufficient to conduct the research inquiry. This could be a very efficient, low cost, and low risk approach to implementing an experiment designed to determine what the impact would be for a device that would ultimately be tied to a smart metering system that will be in place several years hence.

Protocol 10 contains a worksheet that can be used to delineate the key support systems and materials that will be required by the planned experiment. It can also be used to identify systems that could place the experiment at risk and backup plans that can mitigate such risk.

[34] It may be possible to pull a suitable sample from the population that already received new meters using propensity score matching or some other matching scheme. However, since meter deployment is sometimes tied to climate (e.g., meters may be installed in hotter zones first), this may not be possible, in which case it may be necessary to install meters on some customers ahead of normal deployment.

Protocol 10

Please complete the following table. Enter N/A (not applicable) for systems and materials that are not needed for the experiment being designed.

Table 4-7
Key Support Systems and Materials Inventory and Assessment

Description	Fulfillment Plan	Summary of Risks	Alternative Options
Metering			
Meter Data Management			
Billing			
Information Treatments			
Recruitment Tracking			
Recruitment Process			
Marketing Material			
Customer Information/ Education Materials			
Customer Support			
Surveys			
Other			

Budget Planning

Experiments can be expensive and budgeting is an essential part of research design. For large scale experiments involving significant technology applications, research design is often an iterative process in which a preliminary design is produced working through the types of questions contained in Protocols 1 through 10, a preliminary budget is produced and found to exceed the available funds by a large amount. Then, a new plan is produced that fits within the fiscal constraints that exist. Costs will vary significantly depending on the treatments being tested, whether or not interval metering would need to be installed to support the experiment, the ability of existing systems, such as billing, meter data management (MDM), call centers, etc., to support the experiment, and many other factors. Budgeting is not something that requires a formal protocol. Utilities routinely budget for programs and projects , so almost every utility will have established a set of rules and procedures for project budgeting.

Preliminary Schedule

Development of a preliminary schedule is similar to budgeting in that utilities do this type of thing very regularly and have tools and procedures for producing schedules and GANTT charts that show when key deliverables are needed and how various work flows feed into each other. As with budgeting, during the research planning phase of an experimental project, scheduling is often an iterative process in which the needs of the experiment, the drop-dead dates for completion, and typical business practices clash. It may take several iterations of scheduling and research redesign before a feasible schedule, budget, and design are sufficiently aligned.

5
ANALYSIS PROTOCOLS

There are very large number of experimental designs that could be used for feedback research and the equally large number of outputs of potential interest (e.g., change in annual usage, monthly usage, hourly usage or peak demand; change in behavior; customer preferences for various options; insights into how information is used; etc). Therefore, it is impossible to dictate the type of analysis that is most appropriate for every type of research project that can be imagined. Accordingly, the protocols presented herein focus not on how to do the analysis, but on what the analysis should be designed to produce. The analysis protocols presented below and the documentation protocols contained in Section 6 go hand-in-hand and have similar objectives, which include:

- Producing a minimum, common set of outputs for various experiments so more meaningful comparisons can be made across experiments and populations.
- Enhancing understanding of how impacts vary across treatment options, customer characteristics, and other key drivers.
- Ensuring that information is provided that will allow experienced evaluators to assess the validity and accuracy of the findings.
- Enhancing the ability to extrapolate the experimental findings to relevant external customer populations and/or to pool data across utilities.

Methods for Estimating Electricity Consumption Impacts

Depending on the experimental design, estimating load impacts associated with information feedback could be straightforward or complicated. When a simple experimental design, such as the completely randomized design, is used with large sample sizes (possibly optimized to improve statistical precision), fairly conventional statistical techniques can be employed to analyze the resulting data (e.g., difference of means or analysis of variance (ANOVA)). However, because a time series of electricity consumption (e.g., monthly billing data) is almost always available, most of the really simple analysis techniques (i.e., difference of means or difference of differences) are sub-optimal. Using repeated measurement designs will take advantage of the additional information and statistical power that is available from the time series of electricity consumption measurements. This is only a slightly more complicated analysis problem than conventional statistical techniques, and analysis procedures for analyzing the data in this fashion can be found in Stata, SAS and SPSS.

More Complex Designs

When factorial or covariance experimental designs are used, regression modeling techniques[35] will be required to estimate load impacts and to determine how impacts vary across treatments, time, and customer characteristics. This is also the case for most quasi-experiments,

The analysis of load impacts associated with information feedback will most often involve both a time series and a cross-sectional dimension. This type of data is referred to by a variety of names including time series/cross-sectional, panel, longitudinal, and repeated measures data. With such data, analysts are able to account for the effects of two types of omitted variables, fixed effects and time effects, and those that are unobservable or not recorded, leading to unbiased and more robust regression estimates of change in electricity consumption.[36]

Panels

Panel models with both fixed and time effects are strongly recommended for monthly consumption data structures.[37] However, unless a fully randomized assignment is employed, the analysis will need to account for factors that vary across both time and individuals (e.g., electricity prices) and for explanatory variables that are time-invariant (e.g., air conditioning ownership) but interact with time-variant variables (e.g. temperature). These factors, particularly the interaction between air conditioners and weather, are potentially related to the information feedback effect and omitting them from the regression can lead to bias in the treatment effect variable coefficients.

With panel data, each additional observation, whether monthly or hourly, is likely related to previous observations. The auto-correlation between periods needs to be incorporated into the analysis or the confidence bands around the treatment effect may be overstated. A common practice in accounting for auto-correlation is to employ cluster robust-standard errors.[38] This approach generally works for panel data with a limited number of time periods, often referred to as a short panel. For data structures based on monthly consumption data, correcting for clustering in the standard errors for individual customers is often sufficient. For data structures

[35] Use of the term "regression methods" is meant to include methods such as ANCOVA in addition to standard regression methods.

[36] Panel regressions can control for omitted and sometimes unobserved factors that vary across individuals but are fixed over the course of the study (fixed effects – e.g., household size, income, appliance holdings, etc.), and for factors that vary over time but are the same for all customers (time effects, economic conditions). Regression-like models that can be used to analyze panel data include ANOVA, ANCOVA, and MANOVA. These models are similar in that they allow each individual to act as their own control and account for the effects of the fixed, but unmeasured characteristics of each customer.

[37] Fixed and time effect models are more robust to bias than random effects. However, random effect models are generally more efficient. They explain both the variation between and within customers and can reduce the amount of unexplained variance, allowing the detection of smaller effects. Technically, it is possible to test whether a random effects model provides coefficients equivalent to the more bias robust fixed effects model through the Haussmann test. If a random effect model is employed, we highly recommend explicitly employing the Haussmann test and providing readers a side-by-side comparison of effect coefficients for the random and fixed effect models.

[38] Several statistical packages allow for cluster robust standard errors through explicitly defining the cluster group. By clustering, it is assumed that errors are equicorrelated for individual panels. The standard error corrections should be made explicitly in the analysis as several statistical packages assume errors are independent and identically distributed (i.i.d.). The term "robust" is often confused with Huber-White standard errors which account for heteroskedasticity but do not correct for the correlation in individual errors.

with longer time series, such as hourly interval data, auto-correlation in the data structure should be explicitly incorporated into the analysis.[39]

Panel regressions typically calculate the average effect of treatment. For most information feedback programs, this will generally be sufficient. However, in some instances, it will be important to understand the distribution of impacts across participants or the impacts on specific segments of the population. For example, policy makers may be interested in whether the effect differs between low and high income customers or across geographic regions of the utility service territory. Alternatively, the research might be on identifying cost-effective customers and avoiding customers who do not provide response or are cost-ineffective. Overall average impacts can be driven by a small share of customers that provide most of the consumption reduction. As a result, understanding the characteristics that predict high-performance and non-performance can lead to substantial gains in cost-effectiveness.

Panel regressions enable the estimation of impacts for specific segments in the population through the interaction of the treatment effect variable(s) with household characteristics and other relevant variables. However, including too many interactions with the treatment variables complicates the interpretation of results. A key limitation of panel regressions is the inability to produce customer specific impacts.

Isolating and Estimating Customer Effects

Two alternative analysis approaches can provide individual customer effects, which involves tradeoffs in complexity and robustness of results, as follows:

- Hierarchical Linear Models. These are multilevel models than can produce individual customer coefficients and calculate the distribution of impacts. They are primarily applicable to monthly consumption data structures and should be employed with caution. This analysis approach generally assumes random effects, though it is possible to control for time-invariant variables, fixed effects, through group mean centering.[40]
- Individual customer interrupted time series. This approach is only possible with hourly or sub-hourly interval data. It is used to detect individual customer behaviors and effects, relying solely on pre- and post-treatment observations. The approach does not make use of control groups and while it is more prone to bias than panel regressions, it can produce accurate estimates. This approach has been employed with a high degree of success for event-based demand response program evaluations, but it has not been extensively tested with information feedback programs.

While the alternative analyses to characterize customer-specific impacts may be viable, until they are fully tested, it is highly recommended that they be used as supplementary analysis rather than as the basis for estimating program effects. In other words, the best course is to estimate program impacts using a fixed and time effects panel model since it is more robust to bias, makes use of pre- and post-treatment data, and makes use of control group data. If individual customer

[39] For a more detailed discussion on robust standard errors and autocorrelation in short and long panels, please refer to Jeffrey Woolridge's textbook *Econometric Analysis of Cross-section and Panel Data*, pages 274-276, or Cameron and Trevidi's *MicroEconometrics: Methods and Applications*, sections 21.5.2 to 21.5.4.

[40] For further reading see Paul Allison's *Fixed Effects Regression Models.* 2009, and Bryk and Raudenbush's *Hierachical Linear Methods: Applications and Data Analysis Methods.* 2002.

effects are of interest, the above listed approaches can be used to supplement the findings and provide estimates of the distribution of impacts across participants.

Protocol 11: Load Impact Analysis

Establishing the minimum output requirements associated with load impact estimates for information feedback experiments is challenging in light of the significant variation in data, experimental design, and research objectives. For example, for experiments that do not involve interval metering, simply asking that each utility report the average impact by month requires establishing a rule about what constitutes a month given that standard meters are typically read on a read cycle. Should impacts associated with a customer whose meter is read on, say July 10th, be counted as a July impact or a June impact, or should some sort of algorithm be used to split the impact estimate between June and July based on some sort of weather adjusted, day-weighting method?

For experiments that involve interval metering, the level of measure granularity can be constructed uniformly across all research subjects. However, the problem becomes deciding which level of granularity to employ. Should the minimum output requirements include, in addition to monthly kWh impacts: The average impact on each monthly system peak day? The impact during the single hour of system peak for each month? The average impact across, say, a six hour peak period for each month? The average impact for weekdays in each month? The average impact for weekends in each month? The impact in every hour of the year? The challenge stems from having too much information, not too little. Its resolution involves trying to avoid the excess burden of producing outputs that might have interest to others but not to the entity conducting the experiment.

The following protocols apply to the estimation of load impacts stemming from information feedback. As discussed above, depending on what data are collected, there could be a large number of potential load impacts of interest or that could be reported, including the change in annual, monthly, or hourly energy use; the change in average demand for an hour or for a peak period; the change in energy use on specific day types (e.g., weekdays or weekends); etc. Regardless of what output is produced, or through what means, it is essential to know not only what the average impact is for a group of customers, but to also know what the variance is around that average (the standard deviation or standard error of the estimate), and whether or not the estimated impact is statistically significant.

Protocol 11

For analyses based on the difference-in-differences approach using pre- and post-measurements for treatment and control groups, produce the following information for the average customer for each treatment tested:

1. The mean and standard deviation for the treatment and control group for each strata or customer segment delineated in the experiment, and for the group as a whole, for each time period (e.g., annual kWh, monthly kWh, average weekday kWh, peak hour for each monthly system peak day, etc.)[41] for the pre-treatment and treatment time periods.
2. The number of customers included in each calculation in Question 1.
3. The estimated impact and the standard error of the estimated impact for each period, for each strata or customer segment delineated in the experiment, and for the group as a whole and the value of the appropriate measure of statistical significance of the impact (e.g., the t-statistic).
4. For experiments involving stratification of customers, estimate the difference in load impacts across strata and the value of the appropriate measure of statistical significance of any difference across group.
5. For each time period for which a load impact is reported, estimate the cooling degree hours to base 72°F and the heating degree hours to base 65°F.
6. Calculate the average values and standard deviations for all customer characteristics data gathered for each treatment and control group used in the calculations.
7. Calculate whether there are statistically significant differences in all characteristics for which data are gathered between treatment and control groups, and between customers in each stratum.

For analyses involving repeated measures or regression modeling, produce the following:

8. Definitions for all variables used in all estimated regressions, a description of the functional form of the equations, and an explanation of logic underlying inclusion of all variables.[42]
9. A print out of all regression results showing the estimated coefficients, r-squared values, and other relevant statistics provided through standard statistical software packages.
10. The estimated impact and the standard error of the estimated impact for each period, for each strata or customer segment delineated in the experiment, and for the group as a whole and the value of the appropriate measure of statistical significance of the impact (e.g., the t-statistic).
11. The estimated value of load impacts based on long-term normal weather conditions, and the definition of how long-term normal weather is defined.[43]

[41] Also report how each relevant period is defined. For example, for experiments involving kWh meters, how is a month defined in light of the fact that nearly all billing cycles straddle calendar months?

[42] For example,

month I	Dummy variables for month of the year, designed to pick up seasonal effects
dayofweeki	Dummy variables designed to pick up day-of-week effects

[43] This requirement assumes that weather terms are properly included in the regression models, in which case producing estimates is quite straightforward. Weather normalization is not indicated for non-regression based calculations as providing weather normalized estimates is not a trivial extension of the estimation method.

12. For experiments involving stratification of customers, estimate the difference in load impacts across strata and the value of the appropriate measure of statistical significance of any difference across groups.

Protocol 12: Behavioral Change Analysis

As discussed in Section 4, some experiments will be designed to determine whether specific customer behaviors have been changed as a result of exposure to feedback. There are three general categories of behavior change that are of interest in feedback experiments. They are:

1. Behavior associated with the adoption and use of feedback technologies.
2. Behavior related to the timing or magnitude of electricity use for specific uses (e.g., thermostat settings, appliance cycle settings, lighting use, etc.).
3. Replacement/acquisition behavior that may have been influenced by feedback (e.g., shell improvements, more efficient appliances, etc.).

Feedback Adoption Behavior

Adoption behavior relates to the decision to purchase or accept feedback technologies or programs. It will generally be measured by recording the results of efforts to sell or give away feedback mechanisms to representative samples of consumers. Surveys of parties who accept and decline offers of feedback technologies or programs may also be conducted to measure customer attributes that cannot be obtained from utility or publically available information systems to see how these factors influence acceptance. There are well-developed analytical techniques for describing the factors that influence the likelihood that consumers make choices generally referred to as revealed preferences analysis. Examples of econometric techniques that are appropriate for this kind of analysis include: binomial logit models, multinomial logit models, and conjoint analysis. These techniques are very appropriate for analyzing the factors that influence choice behavior and should be applied whenever possible.

Usage Behavior

Two data collection approaches are available for measuring changes in behavior related to the timing and magnitude of electricity consumption – end-use energy consumption measurements and surveys. In theory at least, the most reliable source of information concerning changes in behavior related to the timing and magnitude of electricity consumption can be obtained by measuring the energy consumption of specific end-uses within the households under study.

Data from end-use metering can be analyzed in the same manner as other electricity consumption data, except that the analysis is focused combination of these end-uses at the premise level. Seen in this way, energy consumption by end-use is analyzed using a repeated measures design (measured at hourly, daily, or weekly intervals) in combination with indicator variables representing whether the observations are for treatment or control subjects, before and after the onset of the feedback. This approach to measurement has two major advantages over the survey alternatives that are discussed below. First, it is very accurate and can be made to be statistically very precise. It doesn’t depend on the consumer’s ability to recall potentially subtle changes in the ways that they are using their appliances, changes that may have occurred some time ago. Second, it is unobtrusive, so consumers are unlikely to react to the measurement device by changing their behavior in response to the exposure to the measurement system. The downside of this approach to measuring behavior is that it is expensive.

Another approach to measuring behavior change involves the use of surveys to ask consumers questions designed to determine the changes that may have occurred as a result of exposure to the feedback. The simplest, but least accurate, approach to survey design is a single treatment period or post-treatment survey that asks customers to report changes they may have made in their behavior in the recent past. The accuracy of information obtained in this manner is low for two reasons. First, respondents may be unable to accurately recall changes they made that occurred more than a few days prior to the survey interview and may not be aware of changes that were made by other parties in the household. Second, they may overstate the changes they have made if they believe such changes are socially desirable. This approach to measuring behavior change is not recommended for these reasons.

A more accurate and reliable approach to determining behavioral change through consumer surveys would be to conduct two surveys of treatment and control customers, one before and the other after the treatment was in effect. These surveys should be designed to measure behavior in the recent past, say within the last week or month and should focus on easily answered questions about household energy use behaviors. Examples of such questions are “What is the set point on your thermostat right now?” and “About how many of the rooms in your home that are not currently occupied by people have the lights on right now?” and “Are there any entertainment centers running in rooms in your home right now that are not occupied by anyone?” Of course it is also possible to pose questions about electricity consumption behavior that refer to prior time periods and also to ask questions about the occupant’s perceptions and opinions about energy use in such surveys.

Surveying only the treatment groups can determine whether changes occurred between the pre-treatment and treatment periods for that group, but not whether the changes were caused by the treatment. Other factors could lead to such changes (e.g., headlines about climate change, general information campaigns about the importance of conserving energy, the purchase of a programmable thermostat by a consumer who did not previously have one, etc.). In order to establish causality, it is necessary to obtain the same information on treatment and control customers.

There are two approaches to comparing treatment and control customer behavior. One is to gather pre- and post-treatment data on behavior for both groups. The danger in this approach is that it could cause behavior change (e.g., a Hawthorne affect). In order to eliminate this as a potential source of bias, one could use two treatment and control groups, one each for which the behavioral data (or end-use metering) was obtained and one for which it was not. With this approach, a difference-in-differences calculation could be made for the two control and treatment groups to assess whether any significant behavioral changes occurred between the groups that were the result of the data collection process (that is, whether a Hawthorne effect took place). If it did not, the results from the two control groups could be pooled and compared with the pooled treatment groups.

Self reports of behavior change obtained via surveys are the least reliable basis for indentifying the character of behavior change.

An alternative approach that eliminates any possibility of Hawthorne effects is to simply compare usage patterns between treatment and control customers based on post-test only data. A potential problem with this approach is that it could require much larger sample sizes than would be needed to detect the same difference using the change in behavior within the two groups, as fewer observations are needed to detect a difference in differences than to detect a difference between two groups using only post-test data. Moreover, if there is behavioral decay, then what is observed is neither the initial or final treatment effect.

Replacement/Acquisition Behavior

Replacement/acquisition behavior resulting from exposure to feedback can also be measured using surveys and will probably be incorporated into pre-treatment and post-treatment survey measurements for treatment and control groups. The challenge in measuring replacement/acquisition behavior is that sample sizes that will be used in many feedback experiments (i.e., those involving only a few hundred exposures) are probably going to result in too few replacements or purchases to provide statistically reliable estimates of the impacts of feedback on these behaviors. The only realistic solution to this problem is to increase the sample size in the treatment cells.

Survey data can be analyzed using a variety of statistical techniques. In almost all cases measurements of multiple behaviors or multiple measures of the same behavior will be collected in surveys. Some variation on multivariate analysis of variance (MANOVA) or multivariate regression may be required to estimate treatment effects under these circumstances. Moreover, because multiple comparisons are likely to be carried out in these studies, care should be taken to ensure that statistical tests appropriate for making simultaneous inferences are used in any analysis designed to detect behavior change involving multiple behavior comparisons (e.g., Bonferroni Adjustment).

Protocol 12

1. Are estimates of the rates of adoption of feedback technology or program (the treatments) required as part of the research? If yes:
    a. Describe the data that will be collected to measure the rate of acceptance for each treatment (including any data that must be acquired from third party vendors or surveying).
    b. Describe the statistical techniques that will be used to describe the impacts of customer characteristics (e.g., household lifestyle) and feedback system characteristics (e.g., price) on rate of acceptance.
2. Will end-use metering be used to describe changes in electricity consumption behavior by end-use? If yes:
    a. List the end-uses that will be metered for treatment and control households (e.g. lighting, heating, ventilating, and air conditioning (HVAC), dish washing, etc.).
    b. For each end-use, indicate whether there is an a priori hypothesis concerning how the feedback mechanism may affect the end-use.
    c. Describe the technology that will be used to record and recover end-use measurements.
    d. Generally describe the analysis technique that will be used to identify changes in energy consumption by end-use.
3. Will statistical surveys be used to measure changes in electricity consumption related behavior or appliance acquisition behavior? If yes:

a. Describe how the surveys will be implemented, including:

   i. Whether the surveys will consist of panels (i.e., repeated measurements of the same subject) or cross-sections or both

   ii. How often before and during the test the customers will be contacted

   iii. Measures that are taken in survey design to measure and control for selection effects (due to survey non-response and Hawthorne effects)

b. Generally describe the analysis techniques that will be used to identify changes in consumer perceptions and behavior using the survey data.

Protocol 13: Analysis of Participant Use of Information Feedback

Some experiments will seek to determine what information customers use most, or most successfully, from what they were provided (e.g., usage, cumulative expenditures, progress toward goals, etc.), and how the information and devices are used by consumers to modify their usage and/or purchase decisions.

Examples of questions that might be asked about participant use of information include:

- What screens and displays did the consumer find most informative or use most often?
- What fraction of consumers assess the costs, energy consumption, or environmental consequences of various end-uses (e.g., by turning the devices on and off while observing the change in energy use or expenditure rate through the feedback device)?
- What fraction of consumers used the device to establish monthly goals and to manage progress toward those goals, and how many use the feedback in some other manner?
- How long did it take customers to use learn to use the device?
- Who used the device in the home?
- What did they use it for?
- Did the customers use the device during the entire time it was installed in their home or did they cease using it at some point? If they quit consulting the device, why?
- What information on the device did they consider useless? Why?

Most of these questions, and indeed most questions that can be imagined about the ways people use feedback devices, can be answered using straightforward statistical surveys of treatment customers and using simple descriptive statistics derived from carefully constructed surveys of treatment customers.

Protocol 13

This protocol is only to be completed for studies that are intended to analyze the ways in which consumers use the information that is provided in feedback.

1. Will customers in treatment group(s) be surveyed to study the ways in which consumers used the feedback? If no:

   a. Describe the procedures that will be used to measure the ways in which consumers are using the information provided by the feedback mechanism.

2. If more than one treatment is under study, will a common survey be used in all treatments? If no:
   a. How will consumer responses from the different treatments be compared?
3. Will this survey be carried out during the time that the treatment is taking place? If yes:
   a. What actions will be taken to ensure that this survey does not cause a Hawthorne effect when measuring the effect of the treatment on electricity consumption?
4. How will consumers be selected for the survey?
5. List the research questions the survey is intended to address (e.g. "what screens do consumers find most useful?").
6. List the survey questions that will be asked of consumers to address the research questions (e.g., "Thinking of the times when you have used the (insert feedback device name), what screen do you think provides you with the most useful information?").
7. Describe the statistics that will be used to summarize the responses of customers.

6
DOCUMENTATION PROTOCOLS

Protocol 14: Documentation of Feedback Experiments

In order to facilitate comparison of the results obtained from future feedback experiments, it is highly desirable that certain critical aspects of the research that is undertaken be carefully documented. Protocol 14 is designed to achieve this goal. In general, it sets forth the minimum reporting requirements for feedback experiments.

Protocol 14

This protocol is intended to standardize the reporting of certain critical information that is needed to understand and interpret the results of a feedback experiment. Reports concerning feedback experiments should contain the following information.

1. An executive summary including:
   a. A description of the study objectives.
   b. An overview of the experimental design.
   c. A description of the rate of acceptance of the feedback mechanism during test marketing.
   d. A description of the impacts of the feedback mechanism on energy consumption and/or demand in percentage terms along with an indication of the calculated upper and lower confidence levels associated with reported point estimates.
   e. An executive summary that clearly describes any changes that were observed, if an effort was made to observe changes in consumer behavior or device usage patterns.
2. A description of the feedback mechanisms that were tested along with an explanation of how these mechanisms are supposed to alter consumer behavior – This description should clearly describe the functionality of any equipment that was tested as well as a description of any other experimental factors that were included in the test such as variations, incentives, or other information provided to customers during the tests.
3. A detailed description of the experimental design that was used in the study, including:
   a. All variations in marketing strategies tested or used.
   b. Delivery mechanism/hardware combinations tested.
   c. Any variations in incentives used to enhance recruitment, persistence, and performance.
   d. Variations in other factors (e.g., supplemental information or training) that may have been tested during the study.
4. A detailed description of the sample design, and sampling process, including:
   a. The population of interest.

   b. The sampling frame.
   c. Stratification design if any.
   d. Allocation of initial sample to treatment and control conditions.
   e. Allocation of final recruited sample to treatment and control conditions.
   f. Analysis of selection bias that may have occurred during the sampling process, if any.

5. A description of the historical timeline of the test, including:
   a. Planning phase.
   b. Operational phase.
6. A detailed discussion of the statistical procedures used in the analysis of the data from the study, including:
   a. Detailed specifications of statistical models used to describe experimental outcomes.
   b. Data cleaning procedures used in the study.
   c. Procedures used to control for censoring.
   d. Procedures used to control for selection (if appropriate).
   e. Weighting procedures used and sampling weights.
7. Results reported according to the requirements of the analysis Protocols 11-13.

7
OVERVIEW OF EXAMPLES OF DESIGN PROTOCOL APPLICATIONS

The prior sections describe a series of protocols that can be used to guide the design and analysis of research experiments involving information feedback. Subsequent Sections 8, 9, and 10 contain examples of the application of the design protocols to three case study research projects. This section contains a brief overview of the three case study projects , which were selected to represent feedback categories 2, 5, and 6.

It should be noted that the discussions in Sections 8, 9, and 10 primarily focus on the design protocols, and only referentially to analysis or documentation protocols. Specific examples of the analysis and documentation protocols require detailed data, which are not available for the illustrative examples discussed. The examples contained in the following sections were developed to highlight variation in information feedback options -- different information categories as described in Section 2 -- as well as variation in research objectives – focus on impacts, customer acceptance, understanding behavior, etc. What follows is a synopsis of what each case study entails

Category 2: Enhanced Billing

In the last couple of years, numerous utilities have conducted pilots, or implemented on a large scale, a program that provides information feedback to customers on a monthly or quarterly basis. Many of these Category 2 feedback programs provide energy reports that compare a household's energy use to that of its neighbors along with offering conservation tips. According to a recent study,[44] this service has been introduced at utilities that serve 15% of the U.S. population (but somewhat less are actually receiving the service), including in those in Northern and Southern California, Washington, Minnesota, Illinois, Colorado, and Virginia.

An example of this type of information feedback, which is discussed in Section 8, was included in this report in part because of the industry's widespread interest in it. However, there are also a number of other characteristics about this type of feedback that make it a useful example to include in these protocols, including:

- It is low cost, on a per customer basis, relative to many other types of information feedback.
- It can be made available to customers on an opt-out basis, thus offering the potential to reach a very large share of customers and to generate significant, aggregate energy impacts.
- It focuses attention on important research gaps. In spite of a growing body of research on Category 2 feedback, there remain some important unanswered questions, including whether impacts persist over time (including after the feedback stops) and which elements of the

[44] Hunt Allcott. *Social Norms and Energy Conservation.* MIT and NYU. August 24, 2009. http://www.opower.com/LinkClick.aspx?fileticket=otzFSiC6BJU%3d&tabid=76.

information provided (e.g., neighbor comparisons, conservation tips, etc.) have the greatest influence on changes in energy use.

- It has a low incidence of bias. Concerns about customer acceptance and selection bias are largely averted if the program is universally implemented and widely accepted. This is a relatively low, variable cost option with large economies of scale and scope that can be provided to everyone. While, in theory, customers might "opt out" (e.g., call and ask a utility to stop sending the reports), prior studies reported that the opt-out rate is very small (e.g., less than 2%).[45]
- It is amenable to completely randomized experimental designs. Instead of wide-spread availability, randomly selected control and treatment groups can be constructed and observed to study the impacts of the feedback methods on energy use and underlying behavior.
- Sample sizes are quite large, thus allowing for high degrees of statistical precision even if average impacts are modest.
- The information feedback is based on monthly meter reads, so it can be provided by any utility, with or without an advanced metering system. However, the availability of smart meters would make more detailed consumption profile summaries possible.
- Finally, pre-treatment data already exist on virtually all customers, which further simplify research design and implementation.

A straightforward application of a simple experimental design involving random assignment of households to treatment and control conditions could be used. However, the example presented here is more complicated because of an assumed interest in learning about how the feedback mechanism actually works to change consumer behavior (if indeed it does). The experiment includes multiple treatments in order to isolate the effects of the normative comparisons employed in the feedback messages (comparing usage among customers implies there is some acceptable level defined by a typical customer) from the effect of the energy saving tips that are provided in tandem with these messages. It also seeks to track the underlying behavior of consumers that generate observed savings at various times across the two year treatment period using an innovative panel survey design that avoids Hawthorne effects and other issues that can negatively affect experimental validity.

Category 5: Real-Time, Premise-Level Feedback

The second example is also one that has very widespread interest in the industry. It involves an experimental test of real-time, premise-level information feedback. Characteristics of this Category 5 feedback are quite different from those of Category 2 feedback. They materially affect research design, including: the much higher, per unit cost of real-time feedback devices; the possibility for much larger average savings per customer (which has been reported in some pilots); and the fact that a much smaller percent of customers are likely to accept such devices (even if provided on an opt-out basis, which few utilities would consider). This latter issue means that it is important to understand customer preferences for Category 5 technology options and to address selection issues when designing Category 5 research projects. The high variable cost for devices also means that, in order to keep budgets manageable, the scale of feedback

[45] Ibid.

experiments using these technologies is likely to be considerably smaller than those that can be used in Category 2 studies.

The Category 5 example, contained in Section 9, examines three treatments involving real-time feedback:

1. A simple, low cost IHD that provides very basic data, such as instantaneous and cumulative energy use and corresponding cost
2. An enhanced IHD that provides additional information, such as daily usage profiles, historical comparisons, rate tier alerts (if applicable), CO2 emissions, and also has a goal setting feature
3. Real-time usage information provided to a personal computer, which can display the information in a wide range of custom and user-defined formats and would also allow users to play "what if" games that, for example, estimate how bills would change based on assumed load reductions or load shifting, under different available tariff options. This also allows the consumer to utilize commercial web-based interfaces and calculators like those offered by Google and MSN.

In addition to determining the impact on energy use of the three treatment options, and the relative impacts of one treatment to another, this example also focuses attention on gathering information on customer preferences for the three treatment options. This information can be used to assess the functionality and delivery channels and, when combined with impact data, to identify the option that is likely to be most cost-effective if provided to a broader population.

Selection is a critical issue that must be addressed in this experiment for at least two reasons. First, the "push to PC" treatment is only an option for consumers with personal computers. This fact must be taken into consideration when comparing impacts for the IHD vs. the PC option and when determining customer preferences for different options. Second, any "opt-in" program is likely to attract customers who are different from the average or typical utility customer, especially if the customer has to pay part of the cost. Put another way, compared with the Category 2 example of a randomized design with opt-out delivery (and extremely low incidence of opt-out behavior), the population of participants for a program where consumers must opt-in and pay to do so is likely to be quite different.

The proposed research design is able to address this issue by over-recruiting customers for each treatment (based on direct mail solicitation with clear messaging about "first come, first served" availability of devices) and then assigning some customers to control groups. The research plan also proposes to collect information on customer characteristics (e.g., household type, appliance holdings) as a condition of participation, so the data are available for analysis purposes for both treatment and control customers.

Category 6: Appliance Level, Real-time Feedback

The last example concerns Category 6 feedback, the provision of real-time information for selected appliances and equipment within a household. With current technology, this is a significantly more expensive option than the premise-level feedback devices that produce Category 5 information. Individual devices must be equipped with a mechanism to continuously meter usage and communicate that information to an in-home central repository where it is processed and presented on a device screen or the consumer's PC.

Premise level devices may allow consumers to better understand the relative cost of various end-uses (by turning appliances on or off and observing the change in usage). Thus, any experiment involving appliance level feedback would logically also include a treatment group that would be offered a premise level device (Category 5) so that impacts and cost-effectiveness comparisons can be made between the much more costly, appliance level systems and the relatively low-cost premise level devices. The research plan presented in Section 10 includes two treatments: premise level and appliance level real-time information.

Because of the high cost of the feedback equipment and installation, keeping sample sizes at a minimum is even more important for this research project than for the Category 5 project. The proposed recruitment method for this experiment is typical of what is used for recruiting into clinical trials for health studies, for which advertisements are placed announcing the study and soliciting volunteers that meet pre-screening criteria. Study volunteers are then randomly assigned to treatment and control groups. The research plan calls for monitoring appliance usage for both treatments and control customers in order to estimate the differences in appliance usage that underlie the overall change in energy use at the premise level, and to see if there are differences in appliance-level usage patterns between households that use a premise level device and households that receive appliance level feedback.

8 EXAMPLE APPLICATION OF DESIGN PROTOCOLS FOR CATEGORY 2 INFORMATION FEEDBACK RESEARCH

As discussed in Section 7, in the last couple of years, numerous utilities have implemented pilot or large-scale programs that provide customers with monthly or quarterly information that compares their energy use with that of neighboring or cohort households.[46] These programs typically also provide conservation tips that are tailored to what is presumed to be a household's characteristics.[47] Several studies have recently been published indicating that the average reduction in annual energy use from this type of information feedback is in the 1% to 3% range.[48] Although the average savings are modest relative to total usage, the variable cost for this type of information feedback is low. If these outcomes are repeatable, and the savings are sustainable, feedback can be provided on an opt-out basis to nearly all residential customers, and studies to date indicate that opt-out rates are quite low (1 to 2%).

Given the low cost and possible effectiveness of such programs, they may be capable of producing much larger aggregate (and cost-effective) energy savings than would Category 5 feedback that might produce larger savings per customer (5 to 10%) but only reach a small fraction of the population (5 to 10%). Resolving scale and scope attributes of feedback programs is essential for utilities to each determine which course of action best fits its market circumstances.

This section summarizes an application of the design protocols to a Category 2 feedback option. An example of Category 2 feedback is the service offered by OPOWER™ (formerly Positive Energy), with which a number of utilities are experimenting or implementing.[49] Given that this type of feedback is based on monthly consumption comparisons, it can be implemented by any utility, whether or not they have deployed advanced meters.[50]

[46] *Electricity Use Feedback Pilot and Research Activity*. EPRI, Palo Alto, CA. 2009. 1018979

[47] The tailoring is typically based on data available for all customers rather than on survey data. For example, if it is determined from seasonal usage that a household is likely to have central air conditioning, this customer will receive tips about how to reduce air conditioning energy use.

[48] Hunt Allcott. *Social Norms and Energy Conservation*. MIT and NYU. August 24, 2009, and Ayres, Raseman and Shih. *Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage*. http://ssrn.com/abstract=1434950 .

[49] This discussion is not intended to be a critique or review of the OPOWER services, but is addressed at the class of services for which OPOWER is a prominent provider.

[50] Utilities that provide bundled energy and delivery services through single supplier tariffs have the billing detail needed to provide their customers with detailed self-comparisons and to construct comparisons with other customers. However, some utilities provide only distribution services to customers that purchase the commodity (energy) from a competitive supplier. If the competitive supplier issues a separate bill to consumers, then the utility cannot provide a complete and unified bill comparison without the cooperation of the energy supplier.

Figure 8-1 shows the typical content contained in the reports provided by utilities using the OPOWER Category 2 feedback service.[51] The Social Comparison Module in Figure 8-1 shows a household's energy use relative to its neighbors and to a subset of efficient neighbors, which consists of the lowest 20% of energy users in the comparison group. This comparison also includes emoticons and statements using norms that express social values (e.g., great, good, or below average). The combination of the two forms of normative comparisons may prove to be important, as there is some evidence that customers who find out that they use less than they thought they did (for example, by being told that they use less than even efficient neighbors) may subsequently increase use. Other studies suggest that including value statements indicating that such usage is "great" or "good" can offset this potential boomerang effect.[52]

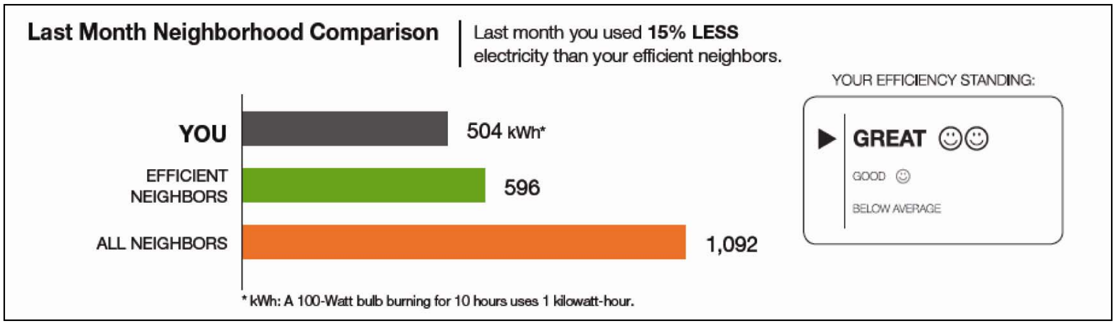


Figure 8-1
OPOWER Social Comparison Module[53]

An additional aspect of this mechanism can be to provide tips on how to conserve energy to help customers use the data to achieve that goal.

The objectives of the research described here go beyond simply estimating the change in energy use resulting from a Category 2 feedback offering. The purpose is to address the following:

1. The feedback experiment has design features and the duration to allow an assessment of long term effects. Specifically, the research will:
    a. Determine whether the effect of the program increases, persists, plateaus, or declines over a reasonable period of time.
    b. Determine whether impacts vary seasonally.
    c. Provide the ability to project the magnitude of expected impacts into future years based on a 24 month time trend.
    d. Measure whether decisions consumers make about appliance purchases are affected by the feedback mechanism.
2. The experiment is designed to quantify the separate impacts of several aspects of the Category 2 feedback designs that are in use today, including:

[51] The references to and illustration of the feedback structure and presentation employed by OPOWER is for illustrative purposes only, and does not constitute an endorsement of OPOWER's products.

[52] See Alcott, 2009 for a brief discussion of these studies and additional references.

[53] Figure taken from Alcott, 2009, Figure 8.1.

   a. The impact of feedback regarding a household's energy use intensity relative to that of its neighbors independent of information about ways to lower energy use (i.e., the effect of the normative messaging alone).

   b. The impact of information about ways households can lower their energy use independent of information about the intensity of a household's energy use relative to that of its neighbors (i.e., the effect of the conservation tips alone).

   c. The combined effects of information about relative energy intensity and helpful hints.

3. The experiment employs a cross-sectional and panel survey design to provide pre-treatment and post-treatment measurements of the energy use related behaviors exhibited by household occupants, as well as awareness of and reactions to messages transmitted to households including messages about relative energy intensity and helpful hints.
4. By withdrawing the feedback stimulus from a subset of the treatment population after one year, while measuring the impact of the stimulus for the remaining treatment population, the study allows for a determination of whether impacts persist after the information is no longer provided.

Protocol 1: Define Treatments and Target Customer Segments

As described in Section 4, the first step in feedback research design is deciding on the treatment options that will be assessed. Protocol 1 contains a table that can be used to describe each treatment option being investigated according to five primary characteristics: information content, format, delivery channel, delivery frequency, and whether or not there are interactive features associated with the treatment (e.g., "what if" analysis capabilities). The target audience (e.g., residential, small commercial, etc.) is also delineated in the table, but only broadly (sub-segments are delineated in Protocol 3).

Table 8-1 summarizes the treatments that are included in this example. The full treatment (i.e., Treatment 1a) is consistent with the typical Category 2 programs that are being considered by a number of utilities today. This treatment includes:

- A comparison of customer energy use with all neighbor.
- A comparison with efficient neighbors, which is defined as the average of those in the lowest quartile of energy users.
- Emoticons and statements depicting social values (e.g., great, good, below average).
- Conservation tips tailored to customer usage patterns and other observable characteristics.

The delivery channel is direct mail, but is typically delivered separately from the monthly bill, both to allow time to complete the analysis and also to increase readership. Studies have reported that consumers are more likely to read a separate mail piece from utilities than read bill inserts. The target customer segment in this example is all residential customers.

In addition to the full treatment, there are three others that are designed to measure the impacts of the basic messages being transmitted and the persistence of the effect of the feedback.

- Treatment 1b has the same content as Treatment 1a, but will only be deployed to participants for 12 months. This will allow for a determination of whether the potential behavior changes resulting from the information feedback will remain after removal of the feedback.
- Treatment 2 will provide the conservation tips only, without the usage comparisons with neighbors. A comparison of energy savings for this group with that of Treatment 1 customers will determine the incremental effect of the comparison feedback.
- Treatment 3 consists of monthly feedback that includes the usage comparison with neighbors for a period of 24 months, but not the conservation tips. A comparison of energy savings for this group with that of the Treatment 1a group will determine the incremental effect of the conservation tips.

Table 8-1
Treatments and Target Customer Segments; Category 2 Feedback Trial

ATTRIBUTE	TREATMENT 1a	TREATMENT 1b	TREATMENT 2	TREATMENT 3
INFORMATION CONTENT				
Description of Treatment	Comparison of customer energy use with all neighbors and efficient (top 20%) neighbors plus emoticons and statements depicting social values (e.g., great, good, below average) – see Figure 8-1 Conservation tips tailored to customer usage patterns and other observable characteristics – see Figure 8-2 Treatment lasts 24 months	Same as 1a except treatment only lasts 12 months	Conservation tips tailored to customer usage patterns and other observable characteristics	Comparison of customer energy use with all neighbors and efficient (top 20%) neighbors plus emoticons and statements depicting social values (e.g., great, good, below average) – see Figure 8-1
INFORMATION FORMAT				
Numerical (toggle through each output)	N/A	N/A	N/A	N/A
Text?	Y	Y	Y	N
Graphical	Y (for normative comparisons – see Figure 8-1)	Y (for normative comparisons – see Figure 8-1)	N	Y – see Figure 8-1 for an example illustration
Other	N	N	N	N

Table 8-1 (continued)
Treatments and Target Customer Segments; Category 2 Feedback Trial

ATTRIBUTE	TREATMENT 1a	TREATMENT 1b	TREATMENT 2	TREATMENT 3
DELIVERY CHANNEL				
Dedicated IHD, Professionally Installed	N/A	N/A	N/A	N/A
Dedicated IHD, Customer Installed	N/A	N/A	N/A	N/A
PCT	N/A	N/A	N/A	N/A
Pushed to PC/TV through USB Device	N/A	N/A	N/A	N/A
Customer Access through Web Portal	N/A	N/A	N/A	N/A
Other	Direct Mail (separate from bill)	Direct Mail (separate from bill)	Direct Mail (separate from bill)	Direct Mail (separate from bill)
DELIVERY FREQUENCY				
Frequency	Monthly	Monthly	Monthly	Monthly
INTERACTIVE FEATURES				
Describe in detail any interactive features provided for each treatment	N/A	N/A	N/A	N/A

Protocol 2: Outcome Variables and Customer Sub-Segments

Protocol 2 poses a series of questions designed to produce an initial list of outcomes that are to be estimated through the research. As indicated in Section 4, outcomes of interest typically include changes in annual and/or monthly energy use, changes in the timing of energy use, changes in consumer behavior (that underlie the change in energy use), understanding the way in which consumers process and use the information provided, and customer acceptance of the treatment being offered. Protocol 2 is reproduced below, with answers and explanations provided following each question.

1. Which of the following outcome variables will the experiment be designed to measure? If the outcomes of interest vary by customer segment, indicate the desired outcomes for each customer segment delineated in Question 1.

   a. Change in annual kWh

      Yes.

b. Change in monthly kWh (designate whether for each month or for selected months)

The change in monthly electricity (kWh) use will be determined for each month over the two-year study period.

c. Change in hourly or sub-hourly kWh (designate sub-hourly intervals) for each hour (or sub-hour) for specific, designated time periods (delineate time periods – e.g., all hours in the year, all-hours in selected months, all hours on selected days within a month such as system peak days, etc.)

No. Hourly data will not be examined in this study.

d. Change in peak demand (kW) for specific, designated times (delineate times – e.g., at time of annual system peak, for each monthly system peak, etc.)

Not measured.

2. Will the experiment seek to identify and quantify the prevalence of the specific types of behavior that change as a result of the treatment? If yes, delineate whether any specific types of behavior are of particular interest (e.g., increase thermostat set point in summer, turn off lights more, etc.).

Yes. The precise list of behaviors for which information will be obtained will be determined at a later date. Examples include:

- *Purchase of one or more energy efficient appliances*[54]
- *Installation of more energy efficient lighting*
- *Minimizing lighting of unoccupied space*
- *Elimination of vampire loads*
- *Use of shorter appliance cycle times (i.e., dish washing or clothes washing)*
- *Substitution of less energy intensive techniques for meeting household needs (e.g., use of line drying some or all of the time)*
- *Changed thermostat settings on central air conditioner or heating system*
- *Installation of energy saving measures (i.e., lighting, hot water management, insulation, etc.)*
- *Other actions that may be suggested in finalizing the design*

3. Will the experiment seek to understand how consumers process and use the information being provided to change their behavior?

Yes. Determining what information is used by consumers and how it is used by them will be a key focus of this study. In part, this will be an outcome of the research design and analysis. The incremental effects of the normative information and the conservation tips will be determined by making comparisons in energy savings across treatment options.

[54] It will be important to develop methods to track purchases of energy efficient equipment influenced by other utility or government programs so as not to erroneously attribute savings to the Category 2 information program. However, since these other programs can be expected to affect the control group as well, any difference between control and treatment groups in the adoption of more energy efficient appliances and energy use practices can be attributed solely to the existence of the feedback mechanism.

In addition , the following information will be gathered through surveys (this list is exemplary, not comprehensive):

- *Memory of receiving reports*
- *Recall of frequency of receiving reports*
- *Reported likelihood of reading reports*
- *Reported recall of the content of the reports*
- *Reported usefulness and rate of adoption of energy efficiency tips*
- *Perception of household energy use relative to that of other similar households (i.e., uptake of the normative information)*
- *Perceived ability to lower energy use*
- *Perceived causes of energy consumption*
- *Attitudes about energy consumption and conservation*

4. Will the experiment seek to understand the key drivers of customer choice associated with various information options and program/marketing methods? If yes, describe the various marketing strategies/offers that will be tested for each information option and market segment.

    No. This is an opt-out program that will be implemented to all consumers, so very low opt-out levels anticipated.

Protocol 3: Delineate Sub-Segment Populations of Interest

Understanding how the change in energy use varies across customer segments can be an important outcome of this research. However, it is unclear at this point exactly how different market segments will react to the various message treatments used in this study. The sample sizes[55] in the treatments should be large enough to support a detailed market segmentation study (i.e., to observe how customers with different characteristics use the information provided) once the data have been collected. There should be a sufficiently large number of customers in nearly all sub-segments of interest to determine with reasonable accuracy what the energy impacts would be and statistical analysis can be done to determine whether energy impacts vary across segments.

Protocol 4: Experimental Design

The next step in research planning is to design the experiment that will be used to determine the impact of the treatment on the outcome variables of interest. The following aspects of this Category 2 experiment make the use of a classical randomized design particularly attractive.

- It is possible to randomly assign customers to treatments and thus to control critical variables under study (i.e., message content and timing of delivery).

[55] The actual sample size calculations are included in Protocol 5 below.

- Experience indicates that while customers may opt out of the experimental conditions, only a small percentage (1 to 2%) are likely to do so – thus there is little reason to be concerned about selection effects.
- Pre-treatment data on energy use exists for nearly all customer premises, the only exception being new customers who don't have a full year of pre-treatment data at the premise they now inhabit.
- The costs of administering the treatment are low relative to many other feedback options (e.g., Categories 5 and 6), allowing for sample sizes that are large enough to estimate one of the primary variables of interest, the change in energy use, with a high degree of precision and to determine impacts for various sub-populations with reasonable precision without using a randomized block design.

In spite of the attractive characteristics of Category 2 outlined, there remain several challenges that must be addressed through careful research design. As indicated in the objectives section, a key goal of this research is to understand the changes in behavior that consumers make and to determine which of the treatment features (e.g., normative comparisons, conservation tips, etc.) are primarily responsible for causing the observed behavioral changes. These objectives will be met through a combination of treatment variations and customer surveys.

It is also important to determine how consumer behavior changes over time. Given the variation in energy use seasonally, one would expect actions taken and impacts to vary seasonally. As such, understanding seasonal variation in behavior will require multiple surveys, as it is difficult for consumers to accurately recall their perceptions and actions when the recall period is very long. In addition, it is important to know how behaviors change over time based on the cumulative effect of the information. Finally, some potential and very important behavior changes, such as purchases of more efficient appliances, occur slowly, typically when such appliances reach the end of their useful life. As such, it is important to capture information over an extended period of time that allows for natural turnover in the stock of appliances, and to capture the information close to the time when such purchases occur.[56]

All of the factors outlined above call for multiple, frequent customer surveys over the course of the study period. However, surveying the same customers repeatedly over the course of two years may lead to selection bias or Hawthorne effects, and most likely both[57]. To manage this

[56] When tracking the change in energy use over time at the individual household level, it is important to also obtain data on exogenous factors that may lead to significant fluctuations in energy use, such as changes in household composition (e.g., a baby born, a teenager moving away to college, etc.), structural changes (e.g., an addition to the house), a change in economic conditions (e.g., loss of a job, entry by a household member into the work force), etc. When estimating impacts for the average customer (e.g., by comparing usage for treatment and control customers), it may not be necessary to track such things since such changes should occur more or less equally among the treatment and control groups.

[57] Repeatedly surveying the same household using the same battery of questions concerning their energy use related behavior and perceptions about energy use could easily cause the survey results obtained from a given population using this technique to be unrepresentative of the majority of study participants who are not exposed to such repeated surveys in two ways. First, consumers exposed to the same surveys questions repeatedly may find the survey experience tedious and tiresome and those who make themselves available and respond to repeated surveys about the same topic at short intervals may not be representative of the wider population under study. Second, exposure to the same battery of questions about energy use related behavior is likely to elevate the awareness on the part of consumers of their own energy related behavior by repeatedly causing them to think about their energy use when they otherwise would not. This process in and of itself may produce behavior change independent of the effect of the treatments.

risk, the research plan calls for selection of multiple control and treatment survey panels, each of which would be surveyed only twice during the two-year study period. Collectively, information from the multiple panels will provide the information that could be produced using a single panel measured repeatedly without risking the validity of the measurements resulting from selection effects or Hawthorne effects.

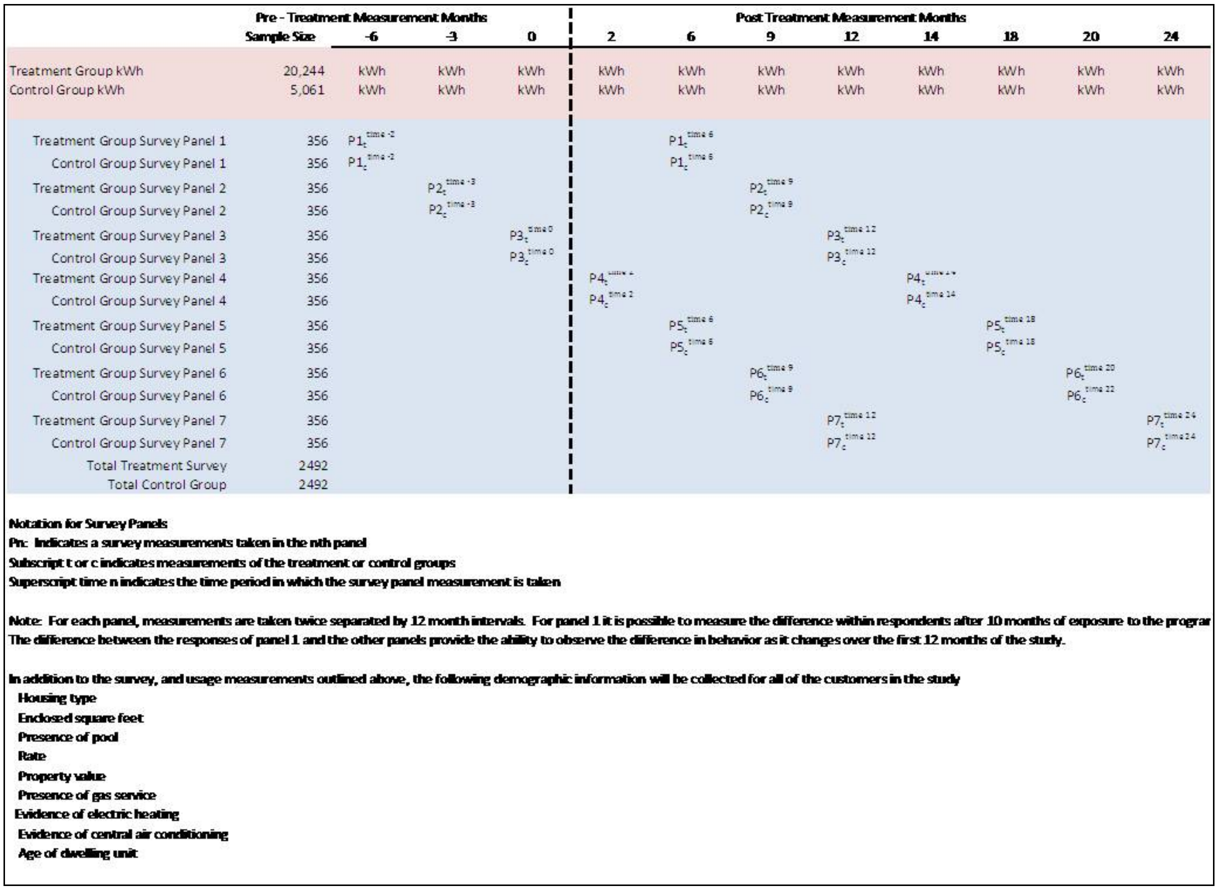
Figure 8-2 provides an overview of the survey approach. Usage measurements will be obtained for all of the 36 months of the study (12 months prior to the start of the feedback and 24 months during which feedback is provided). In addition, seven survey panels will be observed starting in the sixth month prior to the test. Each panel will consist of randomly selected treatment and control households.[58] All panels will be observed twice. The first three panels will be observed during the six months preceding the start of the test. They will be observed again one year later during the same month of the year. The timing of the interviews is very important as reported behavior is likely to vary with season and interviewing at approximately the same time of year provides the best opportunity to do this. It is also notable that the amount of exposure to each treatment for the first three panels varies from six exposures for Panel 1 to 12 exposures for Panel 3.

Measurements for Panels 4 through 7 will commence after the treatment has started – Panel 4 in the second month, Panel 5 in the sixth month, and Panel 6 in the ninth month. The start of each of these panels is timed to coincide exactly with the second time period measurements for Panels 1 through 3. These panels have three purposes. First, they can be used to identify the presence of selection bias and/or Hawthorne effects in the survey measurements taken at the second time periods for preceding panels. In the absence of these effects, the parameter estimates for Panel 1 at time six should be the same as for Panel 5 at time six. If this is not the case, then adjustments must be made for these effects and the differences between responses obtained from Panel 5 at time six (the first time period in which Panel 5 is observed) and Panel 1 at time six (the second time period during which Panel 1 is observed) can be used to develop them.

Secondly, in situations where pre-treatment data are not available, Panel 4 provides a measurement of the differences between treatment and control groups after the first month of the test. Finally, the combination of all seven panels provides measurements of the effects of the treatments on behavior for an extended period (i.e., 24 months from the commencement of the test). The panel observation periods are staggered to provide variation in both season and amount of exposure to the tests.

Collectively, information from the seven panels will allow for measurement of behavioral changes throughout the 24-month study period without introducing the risk of significant Hawthorne effects or selection effects.

[58] Samples sizes for each survey are discussed in the sampling section below.



	Pre - Treatment Measurement Months				Post Treatment Measurement Months							
	Sample Size	-6	-3	0	2	6	9	12	14	18	20	24
Treatment Group kWh	20,244	kWh	kWh	kWh	kWh	kWh	kWh	kWh	kWh	kWh	kWh	kWh
Control Group kWh	5,061	kWh	kWh	kWh	kWh	kWh	kWh	kWh	kWh	kWh	kWh	kWh
Treatment Group Survey Panel 1	356	$P1_t^{\text{time -2}}$				$P1_t^{\text{time 6}}$						
Control Group Survey Panel 1	356	$P1_c^{\text{time -2}}$				$P1_c^{\text{time 6}}$						
Treatment Group Survey Panel 2	356		$P2_t^{\text{time -3}}$				$P2_t^{\text{time 9}}$					
Control Group Survey Panel 2	356		$P2_c^{\text{time -3}}$				$P2_c^{\text{time 9}}$					
Treatment Group Survey Panel 3	356			$P3_t^{\text{time 0}}$				$P3_t^{\text{time 12}}$				
Control Group Survey Panel 3	356			$P3_c^{\text{time 0}}$				$P3_c^{\text{time 12}}$				
Treatment Group Survey Panel 4	356				$P4_t^{\text{time 2}}$				$P4_t^{\text{time 14}}$			
Control Group Survey Panel 4	356				$P4_c^{\text{time 2}}$				$P4_c^{\text{time 14}}$			
Treatment Group Survey Panel 5	356					$P5_t^{\text{time 6}}$				$P5_t^{\text{time 18}}$		
Control Group Survey Panel 5	356					$P5_c^{\text{time 6}}$				$P5_c^{\text{time 18}}$		
Treatment Group Survey Panel 6	356						$P6_t^{\text{time 9}}$				$P6_t^{\text{time 20}}$	
Control Group Survey Panel 6	356						$P6_c^{\text{time 9}}$				$P6_c^{\text{time 22}}$	
Treatment Group Survey Panel 7	356							$P7_t^{\text{time 12}}$				$P7_t^{\text{time 24}}$
Control Group Survey Panel 7	356							$P7_c^{\text{time 12}}$				$P7_c^{\text{time24}}$
Total Treatment Survey	2492											
Total Control Group	2492											

Notation for Survey Panels
Pn: Indicates a survey measurements taken in the nth panel
Subscript t or c indicates measurements of the treatment or control groups
Superscript time n indicates the time period in which the survey panel measurement is taken

Note: For each panel, measurements are taken twice separated by 12 month intervals. For panel 1 it is possible to measure the difference within respondents after 10 months of exposure to the prograr
The difference between the responses of panel 1 and the other panels provide the ability to observe the difference in behavior as it changes over the first 12 months of the study.

In addition to the survey, and usage measurements outlined above, the following demographic information will be collected for all of the customers in the study
Housing type
Enclosed square feet
Presence of pool
Rate
Property value
Presence of gas service
Evidence of electric heating
Evidence of central air conditioning
Age of dwelling unit

Figure 8-2
Planned Survey Approach[59]

Protocol 4 poses a series of questions concerning experimental design, many of which have already been answered in the prior discussion. For completeness, we replicate Protocol 4 below and provide answers to each question.

1. Does the design rely on pre-treatment data?

   Yes. Twelve months of pre-treatment energy use data will be used. In addition, consumer perceptions about energy use and related behavior will be observed for selected panels during the six months prior to the treatment to identify baseline levels for these customer attributes.

2. Do the appropriate data already exist on all relevant customers, or do meters or other equipment need to be installed in order to gather pre-treatment data?

   Twelve months of pre-treatment energy consumption data exist for all customers that have been at their current location for that period of time. There is no need to install additional equipment. Pre-treatment surveys are required to be conducted on three customer panels over the six months prior to the treatment.

3. How long of a pre-treatment period of data collection is required?

[59] The pre-treatment kWh measurements would be used for the full 12-month period prior to the treatment going into effect. This was not shown in the figure in order to fit it onto the page.

See above.

4. Will a control group (or groups) be used in the experiment?

   As depicted in Figure 8-3, there will be an overall control group used to determine energy impacts. From this control group, random samples will be drawn as control groups for the panel surveys carried out during the study.

5. Is it possible to randomly assign observations to treatment and control groups?

   Yes.

6. If random assignment is either inappropriate (e.g., if customers are expected to self-select into the program in the future) or impossible to achieve, how will a suitable control group be selected? Not having a control group is not an option – except under the conditions discussed in Sullivan (2009).

   N/A

7. Using the framework outlined in Section 3, describe treatment(s) and blocks (if any) that will be used during the feedback experiment. This description should be a variation on Figure 3-1, which shows an example of how treatments (and control groups) will be measured for a simple experiment involving two treatments, a control group, and two sampling strata.

   The overall design is depicted in Figure 8-2. It consists of gathering pre-treatment and post-treatment data on a variety of randomized control and treatment groups to determine changes in energy use and changes in behavior underlying changes in energy use.

   The experimental design involves five groups – four treatment groups and a control group. The same measurements and measurement protocols are used in all five groups.

   While this experimental design can be interpreted (and analyzed) as a simple pre-test, post-test design, it is more appropriate to think of it as a factorial design where the levels of the treatment (within a treatment category) vary with exposure (i.e., the number of times the subject has been exposed to the feedback). The data can be analyzed in this fashion or as a simple pre-test post-test design.

Table 8-2
Category 2 Experimental Design

Treatment Group	Pre-Test	Post-Test
Full Treatment 24 months – Treatment 1a Full Treatment 12 months – Treatment 1b Conservation tips only – Treatment 2 Normative Comparison only – Treatment 3 Control Group	Monthly kWh usage for 12 months preceding the test Surveys of representative samples of treatment group members taken at six months, three months, and one month prior to the treatment	Monthly kWh usage for 24 months after the start of the test Follow-up surveys with representative samples from the pre-test one year after the original measurement and at varying intervals after the onset of treatment (six, nine, and 12 months) Additional surveys of representative samples commencing after the start of the treatment

Protocol 5: Sampling

Protocol 5 poses several questions that must be answered in the process of developing the sample to support the experimental design defined in Protocol 4. The questions in this protocol are intended to guide the development of an appropriate sample design and to lead to a reasonably precise description of the sample design and sampling process. The answers to Protocol 5 for this Category 2 research design are as follows.

1. Are the measurements from the experiment to be extrapolated to the broader utility population?

   Yes

   a. If yes, indicate whether the sample will be stratified and what variables will be used in the stratification.

   Sample stratification is not required.

   b. If no, describe the list of customers from which the sampling will be obtained.

   N/A

2. Are precise measurements required for sub-populations of interest?

   No.

   a. If yes, describe the sub-populations for which precise measurements are desired.

   N/A

3. What is the minimum threshold of difference that must be detected by the experiment?

   1% – the effects observed in previous research have ranged from a low of about 1% to a high of 3%. Therefore, to detect meaningful differences between treatment conditions it will be necessary to set detection thresholds at no more than 1%.

4. What is the acceptable amount of sampling error or statistical precision and acceptable level of statistical confidence (i.e., 90%, 95%, and 99%)?

A +/- 1% statistical precision level with 95% confidence – accurate measurements of future cost effectiveness will depend on the precision of this estimate. Again because effect sizes in prior research have been shown to range between 1 and 3%, the minimum sampling precision has been set to 1%.

5. Will customers be randomly assigned to treatment and control conditions or varying levels of factors under study?

   Yes, a random sample of utility customers will be randomly assigned to five experimental groups (four treatment groups plus the control group).

   a. If yes, do you expect customers to select themselves into the treatment condition?

   No; however, it is possible for customers to opt-out of the experiment by requesting that the utility discontinue sending the reports.

   a. If so, how will you correct for this selection process in the analysis and sample weighting?

   On the basis of past experience, the opt-out rate is expected to be in the range of 1-2% and is not expected to have a material effect on the results of the study.

6. If customers will not be randomly assigned to treatment and control conditions or varying levels of factors under study:

   N/A fro a-c below.

   a. Describe the process that will be used to select customers for the treatment group(s).

   b. Describe the process that will be used to select customers for the control group, and explain why this is the best available alternative for creating a non-equivalent control group.

   c. If no control group is used, explain how the change in the outcome variables of interest will be calculated.

7. Describe the sample design that will be used in the study.

   As discussed in Section 4, required sample sizes are a function of the characteristics of the underlying population, experimental design and analysis methods, level of desired precision, and other factors. For purposes of this example, we assume that the characteristics of the residential customer population are consistent with those of utility 1 in Tables 4-4 and 4-5 (as summarized by the coefficient of variation equal to 1.7). Assuming a panel regression approach to the analysis, each group under study must comprise at least 5,061 customers in order to detect a difference of 1% at a 95% confidence level.

   Another very important objective of the study is to identify changes in the actual behaviors of consumers that may result from the different treatment conditions. To eliminate the possibility that these surveys produce Hawthorne effects on electricity consumption, these customers must be in addition to the customers used to estimate impacts on electricity consumption (i.e., in addition to the 5,061 customers described above).

   A wide range of behavioral indicators will be measured in the survey to try to identify how consumers are being affected by the treatments. At the point of designing the study, no prior information exists concerning the likely variation in these perceptions and *behaviors either within survey cross sections or over time. Given the state of the art in studying consumer perceptions and behaviors related to electricity consumption, the proposed survey will be the first attempt to measure such variation. Moreover, concerns about selection effects and Hawthorne effects limit the number of survey observations that can be taken for any given panel of survey observations to two.*

Given the above considerations, it is impossible to derive sample size estimates for survey panels from prior information about variation in measurements of these populations. The repeated measures design (with even two observations per sample point) will undoubtedly significantly improve the power and statistical precision of the measurements, but it is impossible to determine from existing data the magnitude of the improvement in advance of the experiment.

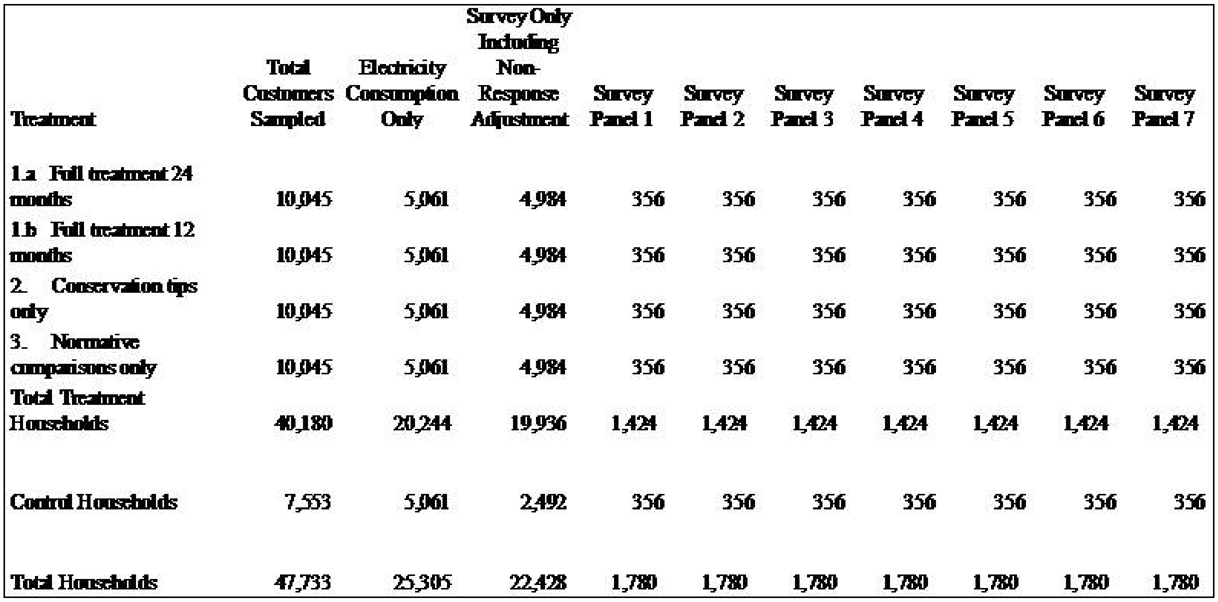
As such, sample sizes for the survey panels were estimated under the assumption that samples should be drawn for each panel to measure proportions derived from the measurements to within plus or minus 6% precision within any panel and time period. The number of observations required to achieve this level of statistical precision is 267. That is, each panel will comprise 267 observations measured at two points in time. Experience suggests that each panel will experience 25% attrition between the first and second measurements. Correspondingly, the sample sizes for the each panel have been inflated to 356 (i.e., 267/.75).

The panels will be analyzed using pooled cross-sections within time series design. In this design, the results from the cross-sections (in this case the panels) are pooled together and indicator variables are used to represent the effects of time and the treatments. In this way the survey measurements from all of the survey panels are aggregated to produce a relatively large sample size for estimating regression models indicating the effects of exposure to the experimental treatments.

Figure 8-3 describes the sample design for this study. All observations for the study will be randomly sampled from the utility's customer records. A total of 5,061 customers will be sampled into each of the treatment groups and the control group to be used to estimate the impacts of the feedback conditions on energy consumption. In order to eliminate the possibility of Hawthorne effects resulting from exposure to the survey questionnaires, these subjects will not be surveyed. An additional 5,061 customers (double the amount of customers required in each survey panel to account for 50% response to the initial panel survey) will be sampled within each group for use in surveying.

From each of the groups under study, seven panels of 712 households will be selected for surveying at different times throughout the course of the project. All of the panels will be surveyed twice and sample sizes have been inflated to reflect the assumption that only 75% of respondents successfully recruited to each panel will respond to the second survey measurement for the panel.

The first three panels will be surveyed prior to commencement of the treatments. The remaining panels will commence after the start of the treatment but will be staggered over the course of the experiment in such a way as to allow measurement of the effect of the treatment on household behavior throughout the course of the study.



Treatment	Total Customers Sampled	Electricity Consumption Only	Survey Only Including Non-Response Adjustment	Survey Panel 1	Survey Panel 2	Survey Panel 3	Survey Panel 4	Survey Panel 5	Survey Panel 6	Survey Panel 7
1.a Full treatment 24 months	10,045	5,061	4,984	356	356	356	356	356	356	356
1.b Full treatment 12 months	10,045	5,061	4,984	356	356	356	356	356	356	356
2. Conservation tips only	10,045	5,061	4,984	356	356	356	356	356	356	356
3. Normative comparisons only	10,045	5,061	4,984	356	356	356	356	356	356	356
Total Treatment Households	40,180	20,244	19,936	1,424	1,424	1,424	1,424	1,424	1,424	1,424
Control Households	7,553	5,061	2,492	356	356	356	356	356	356	356
Total Households	47,733	25,305	22,428	1,780	1,780	1,780	1,780	1,780	1,780	1,780

Figure 8-3
Category 2 Sample Design

While it is impossible to know the statistical power of the survey measurements before conducting the experiment, it will be possible to use the pre-test measurements to study the variability in perceptions and behavior arising from the initial panel measurements and to adjust the sample sizes in remaining panels either upward or downward based on initial measurements of the perceptions and behavior of subject households. In this way the allocation of survey resources across the panels through time can be optimized to achieve the levels of statistical power required to identify changes in important consumer behaviors.

Protocol 6: Recruitment

Given the nature of the information feedback treatments being tested in this experiment, recruitment is not an issue. Both for the experiment and for a full scale program, the information would simply be provided to consumers. Answers to the questions contained in Protocol 6, which covers recruitment, are provided below for this information feedback example.

1. Is the approach to recruitment for a full-scale program that might ultimately be implemented known with certainty?

   Yes.

   a. If yes, does the project timeline allow for experimental recruitment to be done in the same manner as the planned recruitment?

      Consumers are not actually recruited into such a program. They are simply sent the energy usage reports. A small number of customers might call a utility to stop receiving the reports, but prior analysis shows that this number is quite small.

   b. If yes to Question 1a, what is the recruitment approach that will be used (e.g., direct mail, telemarketing, door-to-door, etc.)?

      See above.

   c. If no to Question 1a, what recruitment options fit within the available timeline?

i. What are the potential differences between customers who would be expected to enroll through the long-run recruitment process and customers who would likely enroll through the process that will be used in the experiment?

   N/A

ii. Is it possible to recruit a calibration group using the long-run recruitment approach even if they cannot be enrolled in time to be used in the estimation sample for the load impact analysis?

   N/A

2. Is one of the purposes of the experiment to determine what recruitment process works best and, if so, which options will be studied?

   No.

3. Does the sampling plan involve stratification?

   No.

   a. If so, do data exist that allow for stratification prior to recruitment or does the recruitment process need to gather data on customer characteristics and track enrollment according to these criteria?

   N/A

4. What eligibility criteria, if any, apply to each treatment option?

   No eligibility criteria are being considered.

   a. For each treatment option that has eligibility restrictions, do data already exist that allow for precise targeting of eligible customers?

   N/A

   b. If the answer to Question 4a is no, does the planned recruitment approach allow for eligibility screening to occur and be tracked as part of the recruitment process?[60]

   N/A

5. Taking into consideration the cost of each sample point and any other relevant criteria, how important is it to cut off enrollment as close as possible to the target sample size?

   Unimportant – the cost for each point cost is quite low.

6. If incentives are to be used to enhance subscription, improve persistence, or increase the magnitude of the response to the feedback mechanism, describe the incentives that will be offered and the variations in magnitude of the incentive that will be tested during the experiment.

   N/A

[60] For example, if there was a requirement to have a PC in order to participate with a particular treatment, it is possible to determine whether or not a prospective participant has a PC using telemarketing but not through direct mail. Thus, if direct mail is used for recruitment, it would be necessary to conduct a survey after the fact to determine the enrollment rate among the eligible population (that is, to know of those who did not participate, to distinguish between those who weren’t eligible and those who were eligible but declined to participate).

Protocol 7: Length of Experiment

The length of time that an experiment is run is an important consideration that typically confronts a number of constraints. From a pure research perspective, as a general rule, the longer an experiment is run, the more that can be learned. On the other hand, the incremental cost associated with longer experiments, plus the ever present desire to have answers sooner rather than later, often leads to much shorter time periods than is ideal for research purposes.

In this example, several factors suggest that a two year treatment period and at least a six month pre-treatment period are necessary. One objective of the information feedback is to observe changes in energy use behavior, which varies seasonally. As such, it is important to capture at least one full year of energy use, so that seasonal effects can be observed and the normal variation in seasonal energy use does not lead to erroneous conclusions. As explained in the experimental design section above, it would also be ideal to have a year of pre-treatment data with which to compare the post-treatment behavior on a seasonal basis. For the same reasons, , only three pre-treatment surveys will be conducted over a six month period, as a reasonable tradeoff that would allow for shortening the overall time period by six months compared with the ideal pre-treatment time frame of 12 months. A two year treatment period is included in the research plan in order to assess the cumulative effect of the treatment and, most importantly, to allow for sufficient turnover in the appliance stock to track the impact of the treatments on appliance acquisition behavior.

Answers to the questions posed in Protocol 7 are provided below.

1. Is it possible to run the experiment for at least two years?

   Yes.

   a. If no, how will the persistence of the effect be determined?

   N/A

2. What is the maximum amount of time consumers can be exposed to the feedback mechanism?

   See above (two years).

3. Do pre-treatment data for the relevant variables already exist or must time be allowed to obtain pre-treatment data?

   Yes, pre-treatment data already exist for energy use. The pre-treatment period survey data needed to track changes in behavior do not exist.

   a. If pre-treatment data do not already exist, how long must the pre-treatment period be to support the experimental objectives?

   A six month pre-treatment time period will be used to obtain survey data for detecting behavioral changes.

   b. If pre-treatment data do not already exist, can the experiment be conducted using only post-treatment data, and what adjustments to sample design will be required to employ a post-test-only design?

   Pre-treatment data already exist for energy use and will be collected for consumer behavior.

4. What is the expected amount of time required for consumers to receive and understand the information being provided to them?

   Reports will be provided monthly. Customer understanding of the information will be assessed as part of the study.

5. What is the expected amount of time needed by consumers to implement behavioral changes in response to the information provided?

   The change in behavior over time is an important focus of the research. The impact of information on appliance purchases occurs slowly as the turnover in the appliance stock is slow.

6. How long between the time when a consumer implements a change in behavior and when the feedback associated with that change is likely to be delivered to consumers?

   30 days.

7. What is the minimum amount of time the effect of the feedback mechanism must persist to cost-justify investment on the part of the utility?

   To be determined based on the magnitude of savings estimated through the research.

   a. If the duration of the experiment is shorter than the expected useful life of the measure, how will the determination be made as to whether the effect of the feedback persists long enough to be cost effective?

      N/A

8. Is the feedback mechanism expected to affect consumers' decisions about the energy efficiency or demand responsiveness of new/replacement appliances?

   a. If yes, how will the impact of the feedback mechanism on this behavior be measured?

      The study design includes surveys of approximately 8,000 treatment and control customers over a period of approximately 30 months. Appliance purchase behavior will be tracked for treatment and control customers throughout this period.

9. How much time is needed between when the research plan is completed and approved, and when treatments are in place for experimental participants?

   Given the six-month pre-treatment period needed to obtain survey data on consumer behavior, this question is not particularly pertinent. A more relevant issue is the time required between approval of the plan and when the survey is sent to the first panel group. This time is short, as it only requires the time needed to pull a representative sample of customers and to design and deliver the first survey. This work could be accomplished in a few weeks if needed, although two months would be more comfortable.

10. How much time is required between when the final data are obtained from the experimental observations and when the analysis can be completed?

    Analysis will be done at various times throughout the duration of the project. The final analysis can occur quite quickly, as it will primarily involve comparing results from the final panel surveys and estimating the usage impacts for the last couple of months of the project. This work should be able to be accomplished within a few weeks following receipt of data from the final panel survey and the final billing data.

11. What are the drop-dead dates for when draft and final results from the experiment are needed?

    Estimates will be developed at various points throughout the duration of the project.

Protocols 8 and 9: Data Requirements and Collection Methods

The primary purpose of the protocol pertaining to data requirements and collection methods is to delineate the types of information required, the method that will be used to acquire the information, and how any relevant issues will be addressed that might arise as a result of the data collection process (e.g., Hawthorne effects, survey non-response, etc.). At this stage, it is not possible to fully delineate every specific data element that will be gathered or used to support the analysis. That should be done as part of the design of survey questionnaires or when weather data is being gathered, for example. Rather, the focus here is on delineating the primary types of information that is needed and the sources by which such information will be obtained.

Protocol 8 contains a table that can be used to summarize the types of data that will be gathered during the experiment, the applicable population, the frequency with which it will be gathered over the research period, the method of data collection or source of the data, any issues that might exist with the data or that might arise as a result of the data collection process and how those issues will be addressed. Table 8-3 contains the relevant information for the Category 2 feedback options being addressed in this example.

Table 8-3
Protocol 8: Data Requirements and Collection Methods

Energy use	
Description	Monthly kWh, start and end dates for billing period
Population	All treatment and control customers.
Frequency	Monthly for 12 months prior to the first treatment through the end of the study period, a total time span of 36 months.
Method/Source	Utility MDMS/billing system.
Issues and solutions	A relatively small group of customers will have less than a full year's worth of billing data due to customer churn. Impacts for these customers will be estimated using a comparison between treatment and control customers in the post-treatment period for the subset of customers with this characteristic.
Socio-demographic and appliance data	
Description	Customer characteristics (e.g., income, persons per household, size of house, and appliance holdings).
Population	Treatment and control panel customers.
Frequency	Baseline data obtained from each panel the first time customers are surveyed. Questions about changes in key variables will be asked of the same respondents in the second survey presented to them. Each panel participant will be surveyed twice.
Method/Source	Mail survey (with incentive).
Issues and solutions	Survey non-response. Survey response rates will be maximized through multiple mailings and $2 incentives. Response rates above 60% are common. Will produce more representative sample than using phone. Characteristics of respondents and non-respondents, such as energy use, will be compared to detect any obvious biases.
Energy using behavior	
Description	Data on energy usage behavior (e.g., thermostat settings and habits, number of loads of wash by type (e.g., cold wash, hot wash, etc.), dishwasher usage (number of loads per week, etc.). Same questions asked each survey (that is, customers are not asked to describe how their behavior has changed, just what their behavior is).
Population	Treatment and control panel customers.
Frequency	Data gathered through panel surveys described above.
Method/Source	Same as above.
Issues and solutions	Same as above

Table 8-3 (continued)
Protocol 8: Data Requirements and Collection Methods

Use of information	
Description	Questions about awareness of information being provided, frequency of review of information, which information was used (e.g., normative data, conservation tips, etc.).
Population	Treatment and control panel customers.
Frequency	Data gathered through panel surveys described above.
Method/Source	Same as above.
Issues and solutions	Same as above.
Weather data	
Description	Hourly temperature and humidity for weather stations in close proximity to each customer in control and treatment groups. Will be converted to variables such as cooling and heating degree hours, temperature-humidity index, etc.
Population	All treatment and control customers.
Frequency	Monthly for 12 months prior to first treatment through the end of the study period, a total time span of 36 months.
Method/Source	NOAA and/or other public weather data sources.
Issues and solutions	Careful attention must be paid to geography and micro-climates when assigning customers to weather stations.
Other	
Description	Additional information available from the utility that could be used as explanatory variables in regression models that determine the change in energy use, or to identify high responder customers, and/or to detect non-response bias in surveys, etc., would include such things as prior and future participation in utility sponsored EE and DR programs, tariff, location (for mapping with weather stations and perhaps with publicly available data such as census data), etc. For example, participation in other EE programs, such as appliance rebates, will be an important means of tracking whether the information feedback program or something else influenced future purchases of EE appliances.
Population	All treatment and control customers.
Frequency	Updated on a regular basis (perhaps quarterly) throughout the study period.
Method/Source	Varies – see "Description" section above.
Issues and solutions	None.

As discussed in Section 4, there are advantages to individual utilities and to the industry if each research project gathered a common set of data that would enable comparisons of impacts across experiments and utilities and would support pooling of data across experiments. Protocol 9 lists a common set of data that EPRI recommends be gathered for each experiment. In addition, DOE has issued a minimum set of data requirements, and specific formatting requirements that must be used for any experiment done using funds from the DOE Smart Grid grants. The list of data contained in Protocol 9 includes all of the DOE required variables, but has additional recommended data elements. Protocol 9 asks research planners to identify which of the recommended minimum requirements will not be included as part of the data collection efforts associated with an experiment.

In order to enhance cross-utility comparisons of experimental results or to allow for data pooling across experiments, the following data should be obtained for each experimental subject. Please indicate if any of the data elements are not going to be obtained:

1. A designator indicating the treatment to which the observation was assigned (e.g., Treatment 1, Treatment 2, Control, etc.).
2. For customers in all experiments that do not involve interval metering:
   a. kWh usage for all pre-treatment and treatment billing periods for each participant
   b. Meter read date for each billing period
   c. Monthly electricity bill
   d. Tariff designation
   e. Date that treatment went into effect for all treatment customers
   f. Date customer left experiment for each customer that left before the end of the treatment period
3. For customers in all experiments involving demand-metered customers, in addition to all of the data in Question 1 above:
   a. Monthly peak demand

   Not applicable for this experiment.t
4. For customers in all experiments in which all customers have interval meters:
   a. kWh usage for each hour for the pre-treatment and treatment time periods
   b. Items 1b, 1c, 1d, and 1e

   Not applicable for this experiment.
5. For customers in all experiments, data on the following customer characteristics:
   a. Zip code
   b. Date the customer entered the experiment (treatments or controls)
   c. Date the customer departed from the experiment (treatments or controls)

   NOTE: ALL REMAINING VARIABLES WILL BE GATHERED FOR PANEL SURVEY PARTICIPANTS ONLY

   d. Reason the customer departed from the experiment (treatments or controls)
   e. Presence of central air conditioning
   f. Number of room air conditioners
   g. Presence of electric space heating by type (e.g., base board, heat pumps, etc.)

h. Type of control device for air conditioning and space heating (e.g., standard thermostat, programmable thermostat, etc.)

i. Presence of electric water heating by type (e.g., tank, tankless, etc.)

j. Presence of dishwasher, clothes washer, electric drier, electric cook top, electric oven, electric hot tub/Jacuzzi, swimming pool pump, domestic water pump, Plasma TV

k. Housing type (e.g., single family detached, single family attached, multi-family, etc.)

l. Size of dwelling

m. Number of persons in household by age grouping

n. Annual household income

Protocol 10: Key Support Systems

Another key element of the research design is determining the key systems and materials that will be needed to support an experiment and how those needs will be fulfilled. Protocol 10 contains a table that can be used to identify the key systems and materials that will be needed, delineate the primary fulfillment plan for each, identify any risks that exist and, if relevant, a backup plan. Given the nature of the Category 2 example, most of the needed support can be outsourced and there are few significant risks with fulfillment. Table 8-4 shows how Protocol 10 would be completed for the Category 2 example presented here.

Table 8-4
Key Support Systems and Materials Inventory and Assessment

Description	Fulfillment Plan	Summary of Risks	Alternative Options
Metering	Standard watt hour meters are sufficient.	None	N/A
Meter Data Management	Standard	None	N/A
Billing	Standard	None	N/A
Information Treatments	Monthly reports for each treatment will be required for each participating customer. The neighborhood comparison reports require ongoing analysis of bills of non-participants as well. There is substantial back-office analysis and production required for this that will be outsourced.	Must ensure that the right treatments are sent to each panel.	N/A

Table 8-4 (continued)
Key Support Systems and Materials Inventory and Assessment

Description	Fulfillment Plan	Summary of Risks	Alternative Options
Recruitment Tracking	There is no recruitment required, but it will be necessary to track opt-out customers (those that no longer want the reports to be sent) and normal customer churn among panels (e.g., customers who move).	Must ensure that data is captured and communicated internally and to outsourcing and evaluation contractors.	N/A
Recruitment Process	No recruiting.	N/A	N/A
Marketing Material	No marketing material is needed beyond what is contained in the monthly feedback reports.	N/A	N/A
Customer Information/Education Materials	No information or education materials beyond those that are contained in the monthly feedback reports.	N/A	N/A
Customer Support	The production and delivery of the feedback reports will be outsourced and the outsourcing firm will have a toll free number to answer questions. Utility customer service representatives will be briefed about the project and refer any calls to the outsourcing number.	None	N/A
Surveys	Approximately 12,500 completed surveys among panel customers will be needed over the course of the project. This will be outsourced. Based on a 50% response rate, this will require contacting roughly 28,000 customers.	The response may be higher or lower than assumed, which would affect costs and/or statistical precision.	Include a contingency amount in the budget in case response rates are lower than expected.
Other	N/A	N/A	N/A

Protocol 11: Analysis Plan

As discussed in Section 4, at this stage, the primary purpose of considering how the analysis will be done is to ensure that all of the necessary data are gathered and that sample sizes are large enough to support the required analysis. It is not necessary to specify precisely how all of the analysis will be conducted, but enough thought should be done at this stage to help ensure that nothing has been forgotten that could undercut the entire project or that would be very costly to produce after the fact. As discussed in Section 4, given the wide variety of methods that might be relevant, the analysis plan protocol simply asks designers to describe at a high level what approach or approaches will be used and to ensure that the data necessary to support the analysis have been included in the research plan.

For this Category 2 example, there are a variety of objectives that must be met and the primary analysis approach will vary across these objectives. Very precise load impact estimates can be produced based on a simple "difference-in-differences" approach using monthly usage data for each treatment and control group. This approach will be used to produce a high-level impact assessment for each treatment type that can easily be understood by almost any internal or external audience. Simple statistical tests will be used to assess whether there are any statistically significant differences in the average impact across treatment options.

A more useful analysis approach will involve the use of panel regressions that allow for an assessment of how impacts might vary across customer characteristics (e.g., appliance holdings, socio-demographic characteristics, climate region, etc.). Variables representing customer characteristics can be interacted with treatment effect variables to assess how impacts vary with characteristics. This analysis can be done using the entire treatment sample based on the limited number of variables for which information exists for that larger population (e.g., location, past participation in EE programs, tariff, etc.). The same type of analysis, using the survey data on customer characteristics, can be conducted using the panel data. Another advantage to the regression-based analysis is that it makes it easy to weather normalize the impact estimates or to estimate what the impacts might be for a given set of weather conditions (e.g., a warmer or cooler weather year than what was experienced during the study).

The analysis approach used to determine the change in behavior driven by the treatments will be done using the panel survey data. This analysis will involve a comparison of means between treatment and control customers in each treatment period. For example, referring back to Figure 8-2, the average thermostat setting for treatment and control customers in post-treatment Month 2 for Panel 4 will be compared, using appropriate statistical tests, to determine whether any behavioral changes have occurred early after the treatment goes into effect. The same comparison will be made for, say, Panel 6, to determine whether these changes have become greater, have diminished, or have completely gone away after roughly a year and a half of monthly information has been provided. For the panels and time periods involving the second of the two surveys that each panel will receive, a difference-of-differences comparison can be used to refine the behavioral change estimates.

The analysis of behavioral change will also involve a two-stage modeling approach. In the first stage, changes in behavioral variables between the first and second surveys for each individual will be calculated. These differences will then be used as dependent variables in second stage regressions that relate customer characteristics to changes in behavior. In this manner, one might find, for example, that households with the most significant changes in thermostat settings are also households that participate more in EE programs, or households that have two working members and no children.

Budget

The cost estimates described below are not necessarily indicative of the current market prices of the equipment and services that would be required to actually carry out the study described in this section. They offer as indicative costs that must be considered and the level of detail required for planning.

Design consultant	$50,000 to $75,000
Feedback Cost	$0.84 to $1.4 million per year ($12 to $20 per year per customer for approximately 30,000 treatment customers for two years and 10,000 treatment customers for one year).
Surveys	$.5 million to $.75 million .Based on two times the number of completed surveys indicated in Figure 8-5 (total of about 25,000) spread over roughly 30 months at $20 to $30 per complete.
Analysis	$200,000 to $300,000.
Total Cost	$1.59 million to $2.53 million over roughly 40 months.

The costs above may be more than many utilities would be able to spend. This is typical of research planning, where the preferred design must be reconfigured once budget realities are revealed. There are various ways of reducing the costs, including testing fewer treatments (eliminating the 12 month test, Treatment 1b, for example), lowering the statistical power (thus reducing required sample sizes and survey costs), and measuring smaller and fewer survey panels (which would compromise the ability to detect seasonal effects or persistence).

Schedule

This discussion can be refined but roughly, a 40 month time period overall will be required. Roughly two months are needed on the front end to pull the samples, design the initial survey questionnaire, and initiate arranging for the preparation of the treatment material, either in-house or from a contractor. Then there is the six month pre-treatment period, 24 months of treatment, and about two months on the back end to finalize the analysis and write a report. Most of the analysis will be done prior to the completion of the treatment period with logical analysis points along the way (e.g., six, 12, and 18 months after start of the treatments).

9
EXAMPLE APPLICATION OF DESIGN PROTOCOLS FOR CATEGORY 5 INFORMATION FEEDBACK RESEARCH

This section presents a research plan for an experiment associated with several Category 5 information feedback options. As outlined in Section 7, this experiment has very different objectives and challenges than the Category 2 assessment. One objective is to determine the differential energy impacts associated with each treatment option, but the similarity with the Category 2 example ends there.

Furst, based on prior studies, the average impacts could potentially be significantly greater for Category 5 treatments and, therefore, sample sizes can be much smaller, which is good because the average cost per participating customer is much higher. Second, unlike with Category 2, selection bias is a very important issue to understand and address, as not all customers are eligible for all options and many who are unlikely to take up the offer. Finally, a key area of focus is on understanding the differential acceptance rates by customers among the options being offered.

In order to focus on the issues outlined above, and to keep costs and complexity under control, this example does not focus on understanding the behavior underlying the energy impacts that may arise, or understanding how customers use the information being provided. In short, the focus is on load impacts and acceptance rates, not on characterizing the behavior that underlies these effects.

In order to simplify the example, we assume that the utility in question already has smart meters installed on a sufficiently large number of customers in the target population to fulfill the necessary samples. We also assume that a year's worth of pre-treatment data already exist for these customers. These assumptions materially affect the overall approach that is used, as explained below.

As was true in Section 8, the discussion in the remainder of this section is organized around the 13 research planning protocols presented in Section 4.

Protocol 1: Define Treatments and Target Customer Segments

Table 9-1 summarizes the treatments that are included in this Category 5 feedback example. Three treatment options are to be tested. The first is a simple, low cost IHD that provides very basic data, such as instantaneous and cumulative energy use and expenditures. The second treatment is similar, but provides additional information, such as daily usage profiles, rate tier alerts (if applicable), and CO2 emissions. It also has different formatting capability and a goal setting feature. This device does not allow for custom tailoring of information displays. The third treatment option is to push the usage data to a PC. This can be accomplished, for example, by employing a router with a built-in ZigBee compatible communication device or perhaps through a ZigBee-enabled USB device that would plug into a computer. With this option, software would be provided that has default tables and graphs but that also allows for customization of the information format and content. The software would also allow users to play "what if" games that would, for example, estimate how bills would change based on assumed load reductions or load shifting under different available tariff options.

Table 9-1
Treatments and Target Customer Segments

ATTRIBUTE	TREATMENT 1	TREATMENT 2	TREATMENT 3
INFORMATION CONTENT			
Description of Treatment	Basic IHD (B-IHD) that displays kWh and $	Enhanced IHD (E-IHD) with much more information/toggle through detailed usage screens, projections	Push to PC (could be accomplished through ZigBee device in router or perhaps through USB communicating device).
INFORMATION FORMAT			
Numerical (toggle through each output)	Y	Y	Y
Tabular	N	Y	Y
Graphical	N	Y	Y
Other	N/A	N/A	Can be tailored to consumer's tastes with software provided for PC.
DELIVERY CHANNEL			
Dedicated IHD, Professionally Installed	N	N	N
Dedicated IHD, Customer Installed	Y	Y	N
PCT	N	N	N
Pushed to PC/TV through USB Device	N	N	Y
Customer Access through Web Portal	N	N	N
Other	N/A	N/A	N/A

Table 9-1 (continued)
Treatments and Target Customer Segments

ATTRIBUTE	TREATMENT 1	TREATMENT 2	TREATMENT 3
DELIVERY FREQUENCY			
Frequency	Continuous	Continuous	Continuous
INTERACTIVE FEATURES			
Describe in detail any interactive features provided for each treatment	NONE	NONE	Functionality of PC program makes possible scenario analysis and other interactive features and displays.
CUSTOMER SEGMENTS			
All Residential	Y	Y	Residential with PCs.
Other	N	N	N

Protocol 2: Outcome Variables and Customer Sub-Segments

Protocol 2 poses a series of questions designed to produce an initial list of outcomes that are to be based on the research results. As indicated in Section 4, outcomes of interest could include changes in annual and/or monthly energy use, changes in the timing of energy use, changes in consumer behavior (that underlie the change in energy use), understanding the way in which consumers process and use the information provided, and customer acceptance of the treatment being offered.

The primary focus of this Category 5 experiment is on load impacts – changes in overall energy use at the monthly and annual level as well as in the pattern of energy use hourly – and on customer acceptance among the treatment options. With respect to the latter, the information content and delivery channel (e.g., IHD vs. PC) vary across treatments. While understanding the changes in behavior underlying the impacts is always likely to be of interest, doing so with a high degree of precision would require larger sample sizes and significantly increase costs for what is already an expensive experiment. These issues could be investigated in a later study, based on the single treatment that is preferred by consumers as identified through this project. Information on whether or not consumers used the devices and which device functions were most useful, will be obtained through a combination of surveys for Treatment 1 and Treatment 2 and software tracking and Internet transmission for Treatment 3.

Protocol 2 is reproduced below, with answers and explanations provided following each question.

1. Which of the following outcome variables will the experiment be designed to measure? If the outcomes of interest vary by customer segment, indicate the desired outcomes for each customer segment.

   a. Change in annual kWh

      Yes.

b. Change in monthly kWh (designate whether for each month or for selected months);

   The change in monthly electricity (kWh) use will be determined for each month over the one-year treatment period.

c. Change in hourly or sub-hourly kWh (designate sub-hourly intervals) for each hour (or sub-hour) for specific, designated time periods (delineate time periods, e.g., all hours in the year, all-hours in selected months, all hours on selected days within a month such as system peak days, etc.)

   The change in the pattern of hourly energy use will be investigated. Data for each hour in the year for the 12 months prior to installation of the treatments and the 12 months following installation will be gathered and analyzed.

d. Change in peak demand (kW) for specific, designated times (delineate times, e.g., at time of annual system peak, for each monthly system peak, etc.)

   Yes. Estimates of the average impact for each monthly system peak day during the treatment year will be developed.

2. Will the experiment seek to identify and quantify the prevalence of the specific types of behavior that change as a result of the treatment? If yes, delineate whether any specific types of behavior are of particular interest (e.g., increase thermostat set point in summer, turn off lights more, etc.).

   No.

3. Will the experiment seek to understand how consumers process and use the information being provided to change their behavior?

   Information on whether or not consumers used the devices and which device functions were most useful will be obtained through a combination of surveys for Treatment 1 and Treatment 2 and software tracking and Internet transmission for Treatment 3.

4. Will the experiment seek to understand the key drivers of customer choice associated with various information options and program/marketing methods? If yes, describe the various marketing strategies/offers that will be tested for each information option and market segment.

   Customers will be offered a choice among the three treatment options using the same recruitment method and price point, in order to determine the preferences of customers among the three options.

Protocol 3: Delineate Sub-Segment Populations of Interest

Protocol 3 is used to identify whether or not effects are to be estimated for selected customer segments. In this instance, one of the three treatments can only be implemented by households that have personal computers. Households that own and use PCs may differ from those who don't (e.g., have higher incomes, higher education, etc.) in ways that could be correlated with energy impacts. In order to compare energy impacts across all three treatments, this selection issue must be addressed. If PC ownership were known ahead of time, it would be possible to segment the population into PC and non-PC owning households, and then offer all three treatments to the former and the first two treatments to the latter as a means of determining preferences among eligible households. However, PC ownership is not known a priori so an alternative approach is needed. This approach is outlined in the next section, which eliminates the need for upfront segmentation.

Protocol 4: Experimental Design

As discussed above, there are two primary objectives for this experiment: (1) to determine customer preferences for the treatments offered and (2) to determine the energy impacts associated with each treatment option. Information on device usage will also be gathered. An additional area of interest in determining the impact on usage patterns (e.g., hourly usage, peak demand), not just on monthly or annual energy use. Given the relatively high cost of each of the treatment options, it is important to keep sample sizes as small as possible for those receiving the technologies while still achieving the desired level of precision for the impact estimates. As discussed elsewhere, sample sizes can be dramatically smaller with a pre-test, post-test design compared with using only post-test comparisons between treatment and control customers. In other words, having pre-treatment data is critical.

At the outset of this section, we indicated that this example assumes that a utility already has smart meters installed on a sufficiently large number of target customers to provide the required pre-treatment data. If this is not the case, the approach to customer recruitment and to controlling for selection issues outlined below would not be appropriate. The proposed approach would also work if the objective was to estimate the change in monthly and annual energy use rather than hourly energy use. In this case, it would still be necessary to install smart meters for treatment customers in order to produce real-time information feedback,[61] but they need not be installed on control customers or on all customers during the pre-treatment period. However, if there is a need to know the load impacts at the hourly level, and meters have not yet been installed for a sufficiently long pre-treatment period, a different approach to recruitment and managing selection bias would most likely be needed.

There are three key drivers of the overall experimental design. First is the fact that it is not possible to select random samples of customers with and without PCs beforehand because PC ownership data don't exist. Second is the fact that PC ownership most likely affects treatment selection (as one treatment is only available to PC owners), and PC owners may differ from non-PC owners in ways that affect energy use and demand response. Third is the fact that customers who select any of the options are likely to be different from customers who do not. As such, it is not valid to use a randomly selected control group from the general population as a comparison group for customers who accept any of the treatment options. The most appropriate control group for estimating energy impacts would consist of customers who select an option, but are not placed on the treatment.

The basic research design is summarized in the following steps:

1. Select a random sample of customers from the target population.
2. Customers will be recruited using direct mail (DM) and all customers will initially be offered all three treatment options. In this application, direct mail is preferred over other options such as telemarketing for several reasons: (1) the offer consists of three options that vary in numerous dimensions, which would be hard to convey over the telephone; (2) with modern call screening, it is easier to reach a broad cross section of customers with DM than with

[61] An alternative would be to employ technology such as Blue Line's Power Monitor.

telemarketing; and (3) it could be more awkward telling customers over the telephone who applied than doing it with a postcard. In the DM solicitation, the type of information and functionality associated with each option will be described and customers will be told that they must have a PC (with sufficient power and Internet capability) in order to select Treatment 3. Customers will be asked to mail back a postcard indicating their selection and to check a box on the card indicating whether or not they own a PC. In this manner, it will be possible to later divide respondents into PC and non-PC owning groups, and to estimate preferences and load impacts for households with and without PCs.

3. Importantly, customers will be told that this is a pilot program, that there are only a limited number of devices available and that the devices will be distributed on a first-come basis. Enough DM pieces will be mailed to generate not only a sufficient number of treatment customers for the analysis, but also a sufficient number of control customers for each treatment. The control groups will be comprised of customers who indicate that they want a specific treatment option, but do not get it because the option is over subscribed. For example, suppose that the goal is to get 200 customers to select each treatment option, and the expected acceptance rate for each is 2% of customers who are mailed the recruitment material. Suppose also that the desired size for the control group for each treatment is also 200. Given these assumptions, the goal would be to recruit 600 treatment and 600 control customers, evenly distributed across the three treatments. With an expected response rate of 2%, a utility would need to mail out 60,000 DM pieces to achieve this goal. Under this scheme, half of the respondents would be sent the devices and the other half would be sent a follow up letter indicating that their requests came in too late and the pilot program was fully subscribed. This group of 600 late customers (200 for each treatment) would be used as controls for the 600 customers who receive the treatment devices.[62]

4. While devices could be offered free, in this example, each customer who wants a device will be charged a modest amount (say $25). This puts some "skin in the game" for customers, which may help ensure that they think carefully about the choice they make and that the observed take rates better represent what might be seen if a program were ultimately implemented. It may also better represent how the devices might be marketed in the future, as a utility may not give them away free, but also might not charge full price (especially if energy or peak demand impacts are substantial). While each treatment is likely to have a different cost point in the future, the intent here is to understand customer preferences across the options more than to forecast what take-rates would be under a full scale roll out for a specific device. Thus, it's important to take variation in price across options out of the equation. Once the preferred option is identified (based on insights gained from the pilot), a second investigation could be implemented to determine differential take-rates across various price points and marketing methods.

5. At least two, and perhaps more, recruitment mailings will be implemented in order to gauge differential take rates and manage the recruitment process so that target sample sizes are cost-effectively met and not too many customers are unnecessarily turned away. For example, suppose that the target for each treatment and control group is as specified above (200 each) and the expected acceptance rate based on prior information is 2%. Given these inputs, the

[62] If some customers who respond too late complain about the situation, a utility could actually mail them a device and just exclude them from the treatment and control groups (while recruiting a replacement for them for the control group), could offer them a small financial payment for their trouble, or could offer to send them a device a year later once the pilot is completed.

first mailing might consist of 30,000 DM pieces. From this, we might find, for example, that 100 people want Treatment 1, 150 people want Treatment 2 and 200 people want Treatment 3. As such, subsequent mailings would need to be managed to recruit 300 more Treatment 1 customers (in order to get 100 more devices in place and have 200 matched control customers), 250 more Treatment 2 customers and 200 more Treatment 3 customers. A second mailing of 30,000 would be expected to generate another 200 Treatment 3 customers, thus completing recruitment for this group, and leave Treatment 2 under subscribed by 100 customers and Treatment 1 under subscribed by 200 customers. Given this, a third mailing would be sent, offering only Treatments 1 and 2. The size of the third mailing would be based on the response rates that were generated so far. One approach would be to mail to enough customers to meet the Treatment 1 target, which would lead to over subscription of Treatment 2, in which case, more people would be turned down. Alternatively, the third mailing could be targeted to fill the Treatment 2 cell, and then a fourth mailing, offering only Treatment 1, would be sent at a later date.

The basic approach outlined above generates the desired information on customer preferences for each treatment option and automatically generates a valid control group for each treatment option by identifying customers who want each option but are not placed on the treatment. The required sample sizes for each cell are discussed in the next section. The remainder of this section answers the Protocol 4 questions having to do with experimental design.

1. Does the design rely on pre-treatment data?

   Yes.

2. Do the appropriate data already exist on all relevant customers, or do meters or other equipment need to be installed in order to gather pre-treatment data?

   Twelve months of pre-treatment energy consumption data are assumed to exist for all customers that have been at their current location for that period of time. There is no need to install additional equipment.

3. How long of a pre-treatment period of data collection is required?

   Twelve months of pre-treatment hourly energy use data will be needed in order to complete the analysis and keep sample sizes reasonable.

4. Will a control group (or groups) be used in the experiment?

   A control group will be developed for each of the three treatment options as part of the recruitment process.

5. Is it possible to randomly assign observations to treatment and control groups?

   A random sample will be drawn and used to initiate the recruitment process. Customers will self-select into each treatment based on a marketing approach that initially offers all three treatments to each customer. The recruitment process will be managed so that sample targets are met for each treatment and control group.

6. If random assignment is either inappropriate (e.g., if customers are expected to self-select into the program in the future) or impossible to achieve, how will a suitable control group be selected?

   See discussion above.

7. Using the framework outlined in Section 3, describe treatment(s) and blocks (if any) that will be used during the feedback experiment. This description should be a variation on Figure 3-1, which shows an example of how treatments (and control groups) will be measured for a simple experiment involving two treatments, a control group, and two sampling strata.

   Table 9-2 shows the block diagram for the experimental design. Each treatment option will have a matched control. Separate control and treatment groups will be formed based on PC ownership in order to determine whether impacts vary between PC and non-PC ownership groups.

Table 9-2
Treatments and Target Customer Segments

Treatment	Group Characteristic	Group	Pre-treatment	Treatment
Treatment 1 (Simple IHD)	With PCs	Control	Hourly kWh	Hourly kWh
		Treatment	Hourly kWh	Hourly kWh
	Without PCs	Control	Hourly kWh	Hourly kWh
		Treatment	Hourly kWh	Hourly kWh
Treatment 2 (Additional Functionality)	With PCs	Control	Hourly kWh	Hourly kWh
		Treatment	Hourly kWh	Hourly kWh
	Without PCs	Control	Hourly kWh	Hourly kWh
		Treatment	Hourly kWh	Hourly kWh
Treatment 3 (Push to PC)	With PCs	Control	Hourly kWh	Hourly kWh
		Treatment	Hourly kWh	Hourly kWh
	Without PCs	N/A	N/A	N/A
		N/A	N/A	N/A

Protocol 5: Sampling

Protocol 5 poses several questions that must be answered in the process of developing the sample to support the experimental design defined in Protocol 4. The questions in this protocol are intended to guide the development of an appropriate sample design and to lead to a reasonably precise description of the sample design and sampling process. The answers to Protocol 5 for this Category 5 research design are as follows:

1. Are the measurements from the experiment to be extrapolated to the broader utility population?

   Yes.

   a. If yes, indicate whether the sample will be stratified; and what variables will be used in the stratification.

      The sample drawn from the utility records will not be stratified. However, the observations in the experimental design will be stratified so that ½ of customers in each *treatment are parties with PCs and high speed Internet connections and ½ are parties who do not possess PCs with high speed Internet access. This stratification will be accomplished as the parties respond to the direct mail advertising campaign.*

b. If no, describe the list of customers from which the sampling will be obtained;

N/A

2. Are precise measurements required for sub-populations of interest?

Yes.

a. If yes, describe the sub-populations for which precise measurements are desired.

Precise measurements are required for parties who have PCs with high speed Internet and parties who do not.

3. What is the minimum threshold of difference that must be detected by the experiment?

5%

4. What is the acceptable amount of sampling error or statistical precision and acceptable level of statistical confidence (i.e., 90%, 95%, 99%)?

+/- 5% statistical precision with 95% confidence

5. Will customers be randomly assigned to treatment and control conditions or varying levels of factors under study?

No, parties will be assigned to treatment and control conditions on a first come first served basis. Once the treatment group for a particular experimental cell has been filled, all other parties who respond to the advertising for that group will either be assigned to the control group for that cell or not included in the study at all.

a. If yes, do you expect customers to select themselves into the treatment condition?

N/A

b. If so, how will you correct for this selection process in the analysis and sample weighting?

N/A

6. If customers will not be randomly assigned to treatment and control conditions or varying levels of factors under study:

The order in which parties select themselves into each treatment will be recorded for all parties who indicate that they wish to participate in the study. This variable will be included as a control variable in the analysis.

a. Describe the process that will be used to select customers for the treatment group(s).

See above described customer recruiting process.

b. Describe the process that will be used to select customers for the control group and explain why this is the best available alternative for creating a non-equivalent control group.

The order in which parties respond to the advertising may be related to their interest in the topic and by including a measurement of the order in which parties volunteer this should control this selection effect.

c. If no control group is used, explain how the change in the outcome variables of interest will be calculated.

NA

7. Describe the sample design that will be used in the study.

Table 9-3 describes the sample design for the Category 5 feedback experiment. It is based on a repeated measures design that provides for 12 pre-treatment observations and 12 post-treatment observations for each sample point. The total sample size within each treatment is 400 treatment observations and 400 control observations – with ½ of the observations in each treatment comprised of households with PCs. The sample design also provides for approximately 1,200 households with PCs and 800 households without PCs.

As explained in Section 8, the power of statistical tests is greatly magnified when the outcome variable of interest is measured repeatedly (e.g., monthly electricity consumption). Instead of just one measurement of household electricity consumption before the onset of the treatment and one after, there are actually 12 measures of household electricity consumption before and 12 measures of household electricity consumption after the onset of the treatment – one for each month of the study.

The sample has been designed to estimate the difference in annual electricity consumption for the treatment and control groups to within plus or minus 3% precision with 95% confidence and with 90% statistical power. Practically speaking, this design is capable of detecting at least *a 3% difference in annual electricity consumption 90% of the time for the following comparisons:*

- *Between households with and without IHDs.*
- *Between households with standard IHDs, enhanced IHDs, and those for which information is being pushed to the household PC.*

Table 9-3
Sample Design

Treatment	Group Characteristic	Group	Sample Size
Treatment 1 (Simple IHD)	With PCs	Control	200
		Treatment	200
	Without PCs	Control	200
		Treatment	200
Treatment 2 (Additional Functionality)	With PCs	Control	200
		Treatment	200
	Without PCs	Control	200
		Treatment	200
Treatment 3 (Push to PC)	With PCs	Control	400
		Treatment	400
	Without PCs	N/A	N/A
		N/A	N/A

Protocol 6: Recruitment

The basic approach to customer recruitment was discussed in the experimental design section as these issues are inseparable given the experimental approach. In summary, customers will be recruited using direct mail and, initially, all customers will be offered all three treatment options. Customers will be told that enrollment is on a first come basis and that there is no guarantee that they will be able to obtain a device. Subsequent recruitment waves will be mailed such that sample sizes will be met for the 12 treatment and control group cells shown in Table 9-1. Information on PC ownership will be obtained through the acceptance cards that customers will mail back indicating their technology preference. Customers will be asked to pay a modest price for the information feedback device, with the price being the same for each option so as to assess customer preferences for the attributes of each feedback device, independent of variation in cost. Answers to the questions in Protocol 6 are shown below.

1. Is the approach to recruitment for a full-scale program that might ultimately be implemented known with certainty?

   No. However, there is a reasonably high probability that future recruitment will involve direct mail, perhaps in combination with other recruitment methods, such as telemarketing.

   a. If yes, does the project timeline allow for experimental recruitment to be done in the same manner as the planned recruitment?

      N/A

b. If yes to Question 1a, what is the recruitment approach that will be used (e.g., direct mail, telemarketing, door-to-door, etc.)?

N/A

c. If no to Question 1a, what recruitment options fit within the available timeline?

Direct mail

i. What are the potential differences between customers who would be expected to enroll through the long-run recruitment process and customers who would likely enroll through the process that will be used in the experiment?

N/A (the long run recruitment process is unknown).

ii. Is it possible to recruit a calibration group using the long-run recruitment approach even if they cannot be enrolled in time to be used in the estimation sample for the load impact analysis?

N/A

2. Is one of the purposes of the experiment to determine what recruitment process works best and, if so, which options will be studied?

No.

3. Does the sampling plan involve stratification?

Customers will be stratified into PC and non-PC owning households.

a. If so, do data exist that allow for stratification prior to recruitment or does the recruitment process need to gather data on customer characteristics and track enrollment according to these criteria?

Strata cannot be developed a priori, but data will be collected as part of recruitment to stratify treatment and control groups by PC ownership.

4. What eligibility criteria, if any, apply to each treatment option?

There are no eligibility criteria for Treatments 1 and 2, but customers must own a PC to be eligible for Treatment 3.

a. For each treatment option that has eligibility restrictions, do data already exist that allow for precise targeting of eligible customers?

See above.

b. If the answer to Question 4.a is no, does the planned recruitment approach allow for eligibility screening to occur and be tracked as part of the recruitment process?

See above.

5. Taking into consideration the cost of each sample point and any other relevant criteria, how important is it to cut off enrollment as close as possible to the target sample size?

Cost considerations make it important to limit enrollment in both treatment and control groups to a level very close to the desired sample sizes.

6. If incentives are to be used to enhance subscription, improve persistence, or increase the magnitude of the response to the feedback mechanism, describe the incentives that will be

offered and the variations in magnitude of the incentive that will be tested during the experiment.

N/A

Protocol 7: Length of Experiment

In this example, in one sense, the length of the experiment can be as long as the treatment devices function and as long as monitoring and analysis activities continue. The devices are being sold, for a modest price, to consumers and will not be retrieved. Advanced metering is assumed to be in place so the data needed for energy impact analysis will continue to be available for as long as treatment and control customers remain at the same location. Given these factors, a utility could continue to monitor and evaluate how impacts change over time for as long as the composition of the treatment and control groups does not change significantly due to normal customer churn.

On the other hand, one of the primary objectives of this experiment is to determine which of the three treatment options is likely to produce the best combination of aggregate impacts (e.g., average impacts times take rates) and costs so that the utility can develop a program around that particular option and implement it on a larger scale. Initial customer preferences among the treatment options will be revealed early through the recruitment process. Customer satisfaction with the devices and data concerning which device features are used most frequently will be gathered within the first year by monitoring usage via the Internet for Treatment 3 and through a survey near the end of the first year for the other two devices. These data can be used as input to the selection of the “winning” technology and/or to make modifications in functionality for a device that would be included in the full scale program.

The sample sizes that will be used in this experiment will be too small to determine in any statistical sense what influence the feedback devices will ultimately have on appliance purchases. The slow turnover rate for most appliances means that large samples would be required in order to detect any difference in purchase behavior between treatment and control customers, regardless of how long the experiment is run.

A minimum of one year is needed in order to capture the potential variation in the change in energy use across seasons.

All of the above factors influence the decision to limit the formal treatment period to one year. As indicated above, monitoring and analysis can, and likely would, continue beyond a year, but go-no-go decisions concerning technology choice and formal program development will be made based on analysis after one year of treatment data.

Answers to the questions posed in Protocol 7 are provided below.

1. Is it possible to run the experiment for at least two years?

   Yes, but the primary analysis will be done after one year, in order to make a decision regarding full scale program implementation.

   a. If no, how will the persistence of the effect be determined?

      The treatment devices will not be retrieved at the end of a year. As such, data on energy use for treatment and control customers can continue to be analyzed from normal meter *data as long as the composition of the treatment and control groups remains stable enough to make comparisons meaningful.*

2. What is the maximum amount of time consumers can be exposed to the feedback mechanism?

   A management decision is needed at the end of one year of treatment.

3. Do pre-treatment data for the relevant variables already exist or must time be allowed to obtain pre-treatment data?

   Pre-treatment data already exist.

   a. If pre-treatment data do not already exist, how long must the pre-treatment period be to support the experimental objectives?

   N/A

   b. If pre-treatment data do not already exist, can the experiment be conducted using only post-treatment data, and what adjustments to sample design will be required to employ a post-test-only design?

   N/A

4. What is the expected amount of time required for consumers to receive and understand the information being provided to them?

   Feedback on the rate of energy use and cost is provided in real-time. It is expected that customers will quickly learn the features and functionality of the device. A helpline will be established to assist customers who might have trouble learning how to use the features of the each device.

5. What is the expected amount of time needed by consumers to implement behavioral changes in response to the information provided?

   Changes in usage behavior are likely to occur quickly, but vary seasonally. Changes in purchase behavior, if they exist, occur primarily at the time of the normal turnover in the appliance stock and will not be tracked.

6. How long between the time when a consumer implements a change in behavior and when the feedback associated with that change is likely to be delivered to consumers?

   Feedback on the rate of energy use and cost is in real-time.

7. What is the minimum amount of time the effect of the feedback mechanism must persist to cost-justify investment on the part of the utility?

   To be determined based on the magnitude of savings estimated through the research.

   a. If the duration of the experiment is shorter than the expected useful life of the measure, how will the determination be made as to whether the effect of the feedback persists long enough to be cost effective?

   The trend in change in electricity consumption over the first 12 months of the operation of the devices will be projected to the useful life of the product.

8. Is the feedback mechanism expected to affect consumers' decisions about the energy efficiency or demand responsiveness of new/replacement appliances?

The impact of the feedback mechanism on appliance purchase/replacement is unknown and cannot be observed in this study given the duration of the study and sample sizes involved.

a. If yes, how will the impact of the feedback mechanism on this behavior be measured?

N/A

9. How much time is needed between when the research plan is completed and approved, and when treatments are in place for experimental participants?

The primary determinants of the length of time needed before all treatment devices are in place are the time required to select and acquire the feedback devices, develop the marketing material, and complete recruitment. These three activities must largely be done sequentially, as it would be difficult to develop the marketing material until the devices have been selected (so functionality is known) and recruitment obviously cannot begin until the marketing material is in place. Device selection could easily take two to three months. Marketing material development, approval, and production (of tens of thousands of DM pieces) is likely to take two to three months for a typical utility. The amount of time needed for recruitment will depend on response rates and the number of mailings required. Given the first-come, first-served nature of the recruitment process, customers should be motivated to respond quickly to each mailing. For most DM efforts, few responses are received later than four weeks after mailing. After each mailing, it will take a little time to analyze the data and determine the size and nature of the next mailing in order to hit the sample targets for each treatment and customer segment. As such, recruitment is likely to take two to three months. Given the considerations above, the shortest period between experimental design and treatment implementation is likely to be around six months, and the longest is ten months.

10. How much time is required between when the final data are obtained from the experimental observations and when the analysis can be completed?

Preliminary analysis of load impacts can be done before the final data are produced, so that the final analysis can be completed shortly after the end of the one-year treatment period. A survey will be done near the end of the 12 month period, but analysis of the survey data is straightforward and can be done quickly. It should be possible to complete the final analysis within four weeks of the end of the 12 month treatment period.

11. What are the drop-dead dates for when draft and final results from the experiment are needed?

Two months after the end of the treatment period.

Protocols 8 and 9: Data Requirements and Collection Methods

The data that will be collected as part of this project fall into four broad categories:

1. Usage data on treatment and control customers, which will come from smart meters before and after treatments are put into place
2. Customer characteristics information, which will come from customer surveys
3. Information about customer satisfaction, about whether the device was and is being used, and about which device functions are most useful

4. Revealed preferences among the attributes of the real-time feedback devices that will be offered as part of the project

There are at least three options for collecting data on customer characteristics (e.g., appliance holdings, household size, etc.) from treatment and control customers. One is to have people who want a device fill out a short characteristics questionnaire as part of the selection process. That is, rather than just have respondents check a box on PC ownership (an essential piece of information that must be gathered during recruitment) when they mail back their card indicating they want to purchase a feedback device, a short questionnaire could be included with the recruitment material along with an indication that completing the questionnaire is mandatory if they wish to purchase a device. This approach ensures that such information will be available on all treatment and control customers, is less costly than option 2 (see below), and provides the information at the beginning of the treatment period so that it is available for use in preliminary analysis at any time thereafter. The only possible downside to this approach is that it could reduce participation rates for the experiment. Even if true, this would not materially affect any conclusions that can be drawn from the study, since the primary focus is on relative preferences among the three options, not on predicting what take-rates would be if the program were to be rolled out in the future. As long as including the survey doesn't affect the take-rates for one option relative to another, which is highly unlikely, gathering survey data at this point has little downside and many advantages.

The second option for collecting this information would be to do a survey after the recruitment process is complete. This option has the advantage of getting the data early in the treatment period so it can be used to support preliminary impact analysis and preference modeling, but it is much more costly to do a standalone survey at this point and the survey would be subject to typical non-response bias. Conducting the survey shortly after participants receive the device could also influence the load impact estimates by reminding people about it.

The third option is to collect the characteristics data at the end of the treatment period when a brief survey will be conducted to assess customer satisfaction and to determine whether the device was and is still being used. This approach would have only a small incremental cost, since a survey is going to be done anyway to gather information on device usage. However, it means that the data on customer characteristics will not be available until near the end of the treatment period and, therefore, cannot be used to conduct the preliminary analysis prior to that time.

In light of the above considerations, the first option will be used.

The other primary data to be collected concerns customer use and satisfaction with the devices. We assume here that the devices for Treatment 1 and Treatment 2 do not have the capability of recording information concerning which screens are used or transmitting that information back to a utility through the metering system that the devices communicate with. If such devices do exist and are used in the pilot, that would be both the most accurate and least intrusive way of determining whether the device is being used and, if so, what information is used most. Assuming that this is not possible, the plan calls for a brief survey to be conducted near the end of the treatment period in which such questions would be asked. For Treatment 3, it is assumed that such information can be recorded in the computer software that would be provided as part of the treatment, and could be communicated via the Internet when participants log on and use the software. A brief online survey could also be done with these customers to assess satisfaction and to obtain feedback concerning changes they would like to see in the capabilities of the software.

Table 9-4 summarizes the data requirements and data collection plans for the project.

Table 9-4
Protocol 8: Data Requirements and Collection Methods

Energy Use	
Description	Hourly kWh for 12 months prior to the start of the treatment period, over a 12 month treatment period, and continued monitoring for an indeterminate period thereafter.
Population	All treatment and control customers.
Frequency	Data will be downloaded according to standard meter reading practices (probably daily) and transmitted to the research team monthly.
Method/Source	Utility MDMS/billing system.
Issues and solutions	A relatively small group of customers may have less than a full year's worth of billing data due to customer churn. Impacts for these customers will be estimated using a comparison between treatment and control customers in the post-treatment period for the subset of customers with this characteristic.
Socio-demographic and appliance data	
Description	Customer characteristics (e.g., persons per household, size of house) and appliance holdings.
Population	Treatment and control customers.
Frequency	Obtained through a questionnaire, the completion of which will be a condition of participation.
Method/Source	Survey conducted as part of the recruitment process and is a condition for participation.
Issues and solutions	Approach might reduce response rate to recruitment effort, but should not affect the relative response rates across treatment options, which is most important.
Use of feedback information	
Description	For Treatments 1 and 2, a survey will be conducted at the end of the treatment period to determine if the device had been and was still being used, what features were used most, and what changes they would make in device features if they could. A question about overall satisfaction will be included in this survey. For Treatment 3 customers, the information will be tracked by the software and obtained periodically via the Internet.
Population	All treatment customers
Frequency	A single survey for Treatment 1 and 2 customers. Periodically via the Internet for Treatment 3 customers.
Method/Source	See above.
Issues and solutions	Survey non-response for the Treatment 1 and 2 surveys. Survey methods will be chosen to minimize non-response..

Table 9-4 (continued)
Protocol 8: Data Requirements and Collection Methods

Weather data	
Description	Hourly temperature and humidity for weather stations in close proximity to each customer in control and treatment groups. Will be converted to variables such as cooling and heating degree hours, temperature-humidity index, etc.
Population	All treatment and control customers.
Frequency	Hourly data will be obtained as needed to meet the analysis schedule.
Method/Source	NOAA and/or other public weather data sources.
Issues and solutions	Careful attention must be paid to geography and micro-climates when assigning customers to weather stations.
Other	
Description	Additional information available from the utility that could be used as explanatory variables in regression models that determine the change in energy use, or to identify high responder customers, and/or to detect non-response bias in surveys, etc., would include such things as prior participation in utility sponsored EE and DR programs, tariff, location (for mapping with weather stations and perhaps with publicly available data such as census data), etc.
Population	All treatment and control customers and for a sample of people who were sent marketing material but did not respond
Frequency	Obtained as needed to meet the analysis schedule.
Method/Source	Varies – see “Description” section above.
Issues and solutions	None.

As discussed in Section 4, Protocol 9 lists a common set of data that EPRI recommends be gathered for each experiment.

In order to enhance cross-utility comparisons of experimental results or to allow for data pooling across experiments, the following data should be obtained for each experimental subject. Please indicate if any of the data elements are NOT going to be obtained.

ALL OF THE DATA LISTED BELOW WILL BE OBTAINED FOR TREATMENT AND CONTROL CUSTOMERS.

1. Designator indicating the treatment to which the observation was assigned (e.g., Treatment 1, Treatment 2, Control, etc.)
2. For customers in all experiments that do not involve interval metering:
   a. kWh usage for all pre-treatment and treatment billing periods for each participant
   b. Meter read date for each billing period
   c. Monthly electricity bill
   d. Tariff designation
   e. Date that treatment went into effect for all treatment customers

   f. Date customer left experiment for each customer that left before the end of the treatment period

3. For customers in all experiments involving demand-metered customers, in addition to all of the data in Question 1 above:

   a. Monthly peak demand

4. For customers in all experiments in which all customers have interval meters:

   a. kWh usage for each hour for the pre-treatment and treatment time periods

   b. Items 1b, 1c, 1d, and 1e

5. For customers in all experiments, data on the following customer characteristics:

   a. Zip code

   b. Date the customer entered the experiment (treatments or controls)

   c. Date the customer departed from the experiment (treatments or controls)

   d. Reason the customer departed from the experiment (treatments or controls)

   e. Presence of central air conditioning

   f. Number of room air conditioners

   g. Presence of electric space heating by type (e.g., base board, heat pumps, etc.)

   h. Type of control device for air conditioning and space heating (e.g., standard thermostat, programmable thermostat, etc.)

   i. Presence of electric water heating by type (e.g., tank, tankless, etc.)

   j. Presence of dishwasher, clothes washer, electric drier, electric cook top, electric oven, electric hot tub/Jacuzzi, swimming pool pump, domestic water pump, Plasma TV

   k. Housing type (e.g., single family detached, single family attached, multi-family, etc.)

   l. Size of dwelling

   m. Number of persons in household by age grouping

   n. Annual household income

Protocol 10: Key Support Systems

Protocol 10 contains a table that can be used to identify the key systems and materials that will be needed, delineate the primary fulfillment plan for each, identify any risks that exist and, if relevant, a backup plan. Table 9-5 shows how the Protocol 10 table would be completed for the Category 5 example presented here.

Table 9-5
Key Support Systems and Materials Inventory and Assessment

Description	Fulfillment Plan	Summary of Risks	Alternative Options
Metering	Advanced metering system is assumed to be in place and well functioning.	None	N/A
Meter Data Management	Standard	None	N/A
Billing	Standard	None	N/A
Information Treatments	Three different devices plus a software package, each with the desired functionality, will need to be procured and tested.	There may be technology or communication issues at customer sites, or PC issues, that result in devices that don't work for some customers.	If the number of communication failures is large, additional recruitment could be required to replace the treatment customers who could not use the devices.
Recruitment Tracking	Tracking is an inherent part of the recruitment process, as customers must respond and complete a questionnaire to receive a device.	N/A	N/A
Recruitment Process	Recruitment will be done via direct mail (DM).	Response rates for each option will initially be unknown and there is the risk of over subscription if the initial mail drop is too large.	Limit initial mailing, but this could drag out the length of the recruitment process.
Marketing Material	A well crafted DM piece will be required to describe the purpose and functionality of each treatment option. Depending upon differential response rates, multiple versions may be required (one offering all three to be used until one or more sample cells are full and then another offering the remaining one or two options only).	N/A	N/A

Table 9-5 (continued)
Key Support Systems and Materials Inventory and Assessment

Description	Fulfillment Plan	Summary of Risks	Alternative Options
Customer Information/ Education Materials	Instructions on how to use the treatment options will be provided with the devices and software. It is assumed that suitable material will be obtained from the device suppliers and will not need to be developed by the utility team.	Supplier materials may be found to be unsuitable.	Utility could develop supplemental or replacement materials.
Customer Support	A toll free number will be included with the marketing materials for customers to call and ask questions about the program. A toll free number will be provided to treatment customers so that they can call if they have questions about how to use the device.	Some risk that device suppliers won't meet the service standards that would be ideal from a utility's perspective. A decision will be made as part of the device procurement process concerning whether the device suppliers can or should be used for this or whether a utility technical expert would be better.	Utility staff could be used to man the technical hotline.
Surveys	Survey data on household characteristics and appliance holdings will be gathered as part of the recruitment process. A short survey will be conducted among Treatment 1 and 2 customers near the end of the treatment period to obtain input on whether they used the device, which functions were best, etc. For Treatment 3, that information will be tracked through the software and obtained over the Internet.	Since overall sample sizes for treatment customers will be relatively small, it will be important to maximize response rates to the surveys.	Multiple mailings for surveys, survey completion incentives (e.g., enter into a draw, gift card for those who complete) combined with telephone follow up, can be used to maximize response rates as necessary.
Other	N/A	N/A	N/A

Protocol 11: Analysis Plan

There will be three primary types of analysis done as part of this project: (1) load impact estimation; (2) revealed preference analysis; and (3) descriptive summaries of survey data concerning customer characteristics and appliance holdings, device use and interests in various features and customer satisfaction.

The primary load impact estimation will be done using panel regressions based on pre-treatment and treatment-period hourly data for treatment and control customers. The treatment effect will be estimated using a binary variable representing the presence or absence of the treatment across customers and over time, interacted with other relevant variables that capture variation in energy use across time due to weather and normal usage behavior (depicted by seasonal, monthly, daily, and hourly variables). These interaction terms will determine the extent to which impacts vary seasonally and by time of day.

An important objective of the impact analysis is to determine the relative impacts across the treatment options. This information will be combined with information on differential take-rates from the revealed preference analysis below to determine which option is most likely to produce the highest aggregate impact if offered on a broader scale. The analysis results will also be combined with information on the relative cost of the devices to determine whether the differential costs exceed the differential impacts associated with the option that promises to produce the greatest aggregate change in energy use.

All of the devices will be offered to a large representative sample of utility customers using the marketing procedure that is most likely to be used if the utility decides to go forward with the program. The experiment accurately measures the response of the population as a whole to this marketing effort and the rates at which consumers respond to the advertising can be used to measure their preferences among the alternatives offered. In addition to these gross response rates, it is possible to examine the attractiveness of the various offers to different customer segments using revealed preferences modeling to determine how consumer responses vary for consumers with different impacts. This can be done for all the data in the utility's customer information (e.g., usage, rate categories, credit history, and past history of participation in energy efficiency programs). In addition, it is possible to collect other household descriptive information from third party sources (using address) such as Experian and Nielson. These parties supply household level information such as dwelling size, estimated number of occupants, household income, and lifestyle indicators. Inclusion of these factors in revealed preferences models of consumer preferences may reveal useful information for marketing the choice alternatives in the future and indicate whether certain market segments are attracted to the different technologies.

Summary statistics of the customer use/satisfaction survey will be useful in showing how many people used the devices and for how long, what features were used most and least, what additional features are desired by each group and the overall level of satisfaction.

Budget

The cost estimates described below are not necessarily indicative of the current market prices of the equipment and services that would be required to actually carry out the study described in this section. They are meant as placeholders describing the categories of costs that must be considered and the level of detail required for planning.

Design consultant	\$50,000 to \$75,000
Devices	\$130,000 (400 Standard IHDs @ \$75 + 400 Enhanced IHDs @ \$125 + 400 routers @ \$125 Device)
Shipping	\$60,000 (1,200 @ \$5.00)
Recruitment	\$480,000 (2400 responses required, 2.5% response rate @\$5 per piece)

Surveys	$160,000 (1,600 survey completes @ $100)
Analysis	$100,000 to $200,000
Total cost	$1.00M to $1.11M
Offsetting revenue	$30,000

Schedule

Overall, the project schedule will require six to nine months from the time of project approval until when treatments are sent out, 12 months for the treatment period, and one to two months to complete the analysis and produce a report.

10
EXAMPLE APPLICATION OF DESIGN PROTOCOLS FOR CATEGORY 6 INFORMATION FEEDBACK RESEARCH

Category 6 feedback technologies provide this information at the end-use level in real-time, which allows consumers to see how much they are using for what purposes up to the minute they are viewing the screen.

A more disaggregate view of electricity consumption history and rate of use may induce greater change in electricity consumption related behavior. The disaggregate view provides information about the potential (energy, cost, or CO2) savings that are available from curtailing or rescheduling different kinds of end-uses. In essence, it supplies more detailed and therefore presumably more actionable information that can be used by consumers to decide what to do to lower their electricity consumption or cost. As reviewed by EPRI[63], some studies have suggested that appliance-level feedback may be effective in encouraging conservation, although empirical questions remain regarding its cost-effectiveness, particularly considering its higher potential cost to implement.

Real-time premise level information about electricity consumption can be presented to consumers at a relatively low cost (i.e., between $75 and $200 per premise depending on system design). In the grand scheme of energy efficiency measures, this is not a very large investment per household. Studies conducted over the past two years have found a fairly wide range of changes in electricity consumption resulting from providing real-time feedback using IHDs (i.e., 0 to over 18%). At the upper end of the range, it is likely to be a very cost effective strategy for improving energy efficiency but at the lower end of the range it most likely is not.

Real-time end-use level feedback is more expensive than real-time premise level information because of the requirement to measure and manage information about loads at the end-use or service panel levels. The equipment required to acquire and manage end-use load data presently costs upwards of $2,500 per premise, although much lower priced systems are now being piloted. No one really can say at this point what the large scale cost of the real-time end-use measurement systems will be in the future. Moreover, whether or not it would be cost justified depends largely on the energy efficiency improvements that can be gained from providing premise level feedback – also not known precisely at this point. If gains achievable from premise level feedback turn out to be nil (or even modest), and end-use level feedback produces much larger improvements, then scale economies will be realized and end-use level feedback is likely to become the most attractive approach in the long run, at least from a utility provision perspective. Some customers may find that they can realize benefits that exceed the cost and invest themselves.

[63] *Residential Electricity Use Feedback: A Research Synthesis and Economic Framework*. EPRI, Palo Alto, CA: 2009. 1016844.

The example chosen to illustrate the use of the protocols for a Category 6 research project is a test to determine the incremental benefit (e.g., the change in energy use, if any) attributable to providing end-use level information over and above the provision of only premise level information on an IHD.

Protocol 1: Define Treatments and Target Customer Segments

Table 10-1 summarizes the treatments that are included in this section's design exercise. There are two treatments. In Treatment 1, households will be given enhanced IHDs that present them with information at the premise level for a period of 12 months. In Treatment 2, households will be given end-use level feedback information, in addition to premise level information, for a period of 12 months. Although there certainly are IHD devices that can be installed by consumers, all households in the study, including control households (to collect comparable data, even though it is not displayed to them), will require installation of identical end-use monitoring equipment. As such, professional installation of devices will be needed for all treatment and control customers.

Table 10-1
Treatments and Target Customer Segments

ATTRIBUTE	TREATMENT 1 Enhanced IHD	TREATMENT 2 Enhanced IHD with End-Use Level Information
INFORMATION CONTENT		
Description of Treatment	Enhanced IHD (E-IHD) providing tabular and graphic displays of historical usage, electricity cost, and other metrics in real-time in a device situated within the dwelling.	Same as E-IHD but provides detailed information about electricity consumption by end-use.
INFORMATION FORMAT		
Numerical (toggle through each output)	Y	Y
Tabular	Y	Y
Graphical	Y	Y
Other	N	N
DELIVERY CHANNEL		
Dedicated IHD, Professionally Installed	Y (IHD communicates with Smart Meter at 4 second intervals using Zigbee).	Same as Treatment 1.
Dedicated IHD, Customer Installed	N	N
PCT	N	N
Pushed to PC/TV through USB Device	N	N
Customer Access through Web Portal	N/A	N/A
Other	N/A	N/A
DELIVERY FREQUENCY		
Frequency	Real-time.	Real-time.
INTERACTIVE FEATURES		
Describe in detail any interactive features provided for each treatment	N/A	N/A
CUSTOMER SEGMENTS		
All Residential	Y	Y
Other	N	N

Protocol 2: Outcome Variables and Customer Sub-Segments

Protocol 2 poses a series of questions designed to produce an initial list of outcomes that are to be estimated through the research. Protocol 2 is reproduced below, with answers and explanations provided following each question.

1. Which of the following outcome variables will the experiment be designed to measure? If the outcomes of interest vary by customer segment, indicate the desired outcomes for each customer segment delineated in question 1.

   a. Change in annual kWh

      Yes.

   b. Change in monthly kWh (designate whether for each month or for selected months)

      Yes

   c. Change in hourly or sub-hourly kWh (designate sub-hourly intervals) for each hour (or sub-hour) for specific, designated time periods (delineate time periods, e.g., all hours in the year, all-hours in selected months, all hours on selected days within a month such as system peak days, etc.)

      Yes. Hourly data on electricity consumption by end-use will be analyzed for treatment and control groups.

   d. Change in peak demand (kW) for specific, designated times (delineate times, e.g., at time of annual system peak, for each monthly system peak, etc.)

      Yes.

2. Will the experiment seek to identify and quantify the prevalence of the specific types of behavior that change as a result of the treatment? If yes, delineate whether any specific types of behavior are of particular interest (e.g., increase thermostat set point in summer, turn off lights more, etc.).

   Yes. Behavior changes in this study will be identified in two ways. First, the timing and magnitude of electricity consumption by end-use will be compared for treatment and control groups. Second, household occupants will be interviewed in both treatment groups after 90 days of exposure to both treatments. Measures of behavior change will include:

   - *Changes in the quantities of electricity used in various end-uses such as lighting, heating, water heating, air conditioning and entertainment center*
   - *Changes in the timing of electricity use for the above end-uses*
   - *Self reported changes in electricity consumption related behavior including:*
     - *Elimination of vampire loads*
     - *Use of shorter appliance cycle times (i.e., dish washing or clothes washing)*
     - *Substitution of less energy intensive techniques for meeting household needs (e.g., use of line drying some or all of the time)*
     - *Changed thermostat settings on central air conditioner or heating system*

- ➢ *Installation of energy saving measures (i.e., lighting, hot water management, insulation, etc.) other actions that may be suggested in finalizing the design*

3. Will the experiment seek to understand how consumers process and use the information being provided?

   Yes, at the conclusion of the experiment interviewers will discuss the ways in which occupants used the systems that have been installed in their homes. Topics that will be discussed during the interviews will include:

   - *Whether they are still using the system*
   - *If they are not, when they stopped using it*
   - *How often they have used the display*
   - *What functions they found most useful and not useful*
   - *Who else in the home might have used the display*
   - *What they believe about their electricity consumption based on what they have seen from the different display screens*

4. Will the experiment seek to understand the key drivers of customer choice associated with various information options and program/marketing methods? If yes, describe the various marketing strategies/offers that will be tested for each information option and market segment.

   Yes, during post-test interviews, customers will be asked their reasons for preferring the information found in some screens over others.

Protocol 3: Delineate Sub-Segment Populations of Interest

No effort will be made in this study to observe effects in different market segments.

Protocol 4: Experimental Design

The next step in research planning is to design the experiment that will be used to determine the impact of the treatment on the outcome variables of interest.

The purpose of this experiment is to measure the incremental impact (if any) that results from providing feedback based on end-use level electricity consumption information to consumers above and beyond the impact that can be achieved by providing feedback based on premise level information. The technology required to provide feedback based on end-use level consumption information is available today, but it can be expensive. Therefore, it is important to carefully control recruiting to ensure that only a limited number of parties are included in the test. The salient research question is: Does the added expense produce enough additional impact to be warranted?

While intuitively compelling, it is not advisable to conduct this test on a representative sample of customers. First, given the cost of these installations, utilities are unlikely to offer them to customers who do not want them. Hence, the population of interest is comprised of customers who agree to have the device installed. Second, end-use level IHDs are not widely available in the mass market at this time and the fraction of customers who are likely to respond to direct mail advertising is probably quite small. Given the requirement to have professional installation of end-use monitoring equipment and other study requirements, response rates are likely to be even lower than for the Category 5 project. Finally, little is known about consumers who might be interested in these devices and therefore it will be difficult to target them for purposes of mailing or telemarketing. Consequently, the cost of conventional marketing (i.e., mailing or telemarketing) at this early stage of the market is likely to be quite high.

An approach that may be more effective and much less costly under the circumstances is one that is commonly used in clinical trials. In clinical trials, where patients cannot be recruited through physician networks (because their diseases are rare, undiagnosed, or untreated), recruitment is typically done using an advertising campaign on radio, television, and in the newspaper (in a limited media market) that solicits study participants. The advertising describes the requirements for participation in the study, the study benefits, and the actions interested parties must take to find out if they are qualified to participate.

In this case, these ads would indicate that a company (probably the research contractor and not the utility) is testing a new device that is designed to help household occupants understand and control their electricity consumption. The advertising might say that the device can be used to identify opportunities for saving energy in their home. The value of the product could be described (say $200). In return for participating in the study, participants will receive both the device and the sum of $200 in return for allowing the researchers to install some special monitoring equipment in their home and be interviewed at the end of the study period of 12 months. There may also be other self-reporting requirements.

During the qualifying interview, background information is collected on the participating household that may be useful during subsequent statistical analysis. Qualified respondents are then randomly assigned to treatment and control conditions and the effects of the experimental variables are observed.

The magnitude of changes in electricity consumption that might arise from supplying information about premise level electricity consumption is unknown, as is the impact of providing more detailed information like electricity consumption by end-use. However, based on prior research, it is reasonable to expect these effects to be subtle (i.e., 2-8%). Moreover, there is a great deal of variation in monthly electricity consumption from household to household arising from differences in building design, equipment, occupancy patterns, and other behavioral considerations. For these reasons, it is important to employ an experimental design that provides for:

- Random assignment of those that pass the screen and agree to participate in the treatment and control conditions.
- Pre-treatment measurements of monthly electricity consumption for a period of at least 12 months prior to commencement of the treatments (to detect changes in monthly electricity consumption).
- Measurement of electricity consumption by end-use for treatment and control customers before and during exposure to the feedback variations.

Protocol 4 poses a series of questions concerning experimental design, many of which have already been answered in the prior discussion. For completeness, we replicate Protocol 4 below and provide answers to each question.

1. Does the design rely on pre-treatment data?

   Yes. Twelve months of pre-treatment electricity consumption data will be collected for all customers in the study. In addition, measurements of electricity consumption by end-use will be accumulated for 60 days prior to exposure to the start of the feedback exposure period.

2. Do the appropriate data already exist on all relevant customers, or do meters or other equipment need to be installed in order to gather pre-treatment data?

   Twelve months of pre-treatment electricity consumption data exist for all customers that have been at their current location for that period of time. Pre-existing data on electricity consumption by end-use does not exist, so measurement equipment will be installed 60 days prior to the start of the exposure period to measure pre-treatment electricity consumption by end-use.

3. How long of a pre-treatment period of data collection is required?

   See above.

4. Will a control group (or groups) be used in the experiment?

   Yes.

5. Is it possible to randomly assign observations to treatment and control groups?

   Yes, random assignment to treatment and control cells will be made among the study volunteers.

6. If random assignment is either inappropriate (e.g., if customers are expected to self-select into the program in the future) or impossible to achieve, how will a suitable control group be selected? Not having a control group is not an option – except under the conditions discussed in Sullivan (2009)

   N/A

7. Using the framework outlined in Section 3, describe treatment(s) and blocks (if any) that will be used during the feedback experiment. This description should be a variation on Figure 3-1, which shows an example of how treatments (and control groups) will be measured for a simple experiment involving two treatments, a control group, and two sampling strata.

   The overall design is depicted in Table 10-2. It consists of gathering pre-treatment and post-treatment data on randomized control and treatment groups to determine changes in electricity consumption.

   The experimental design involves three groups – two treatment groups and a control group. The same measurements and measurement protocols are used in all three groups.

   This experimental design is a simple pre-test, post-test design with repeated measures, two treatments, and a control group.

Table 10-2
Category 6 Experimental Design

Treatment Group	Pre-Test	Post-Test
Treatment 1 – Enhanced IHD Treatment 2 – IHD with end-use level electricity consumption displays Control Group	Monthly kWh usage for 12 months preceding the test. Hourly kWh per hour by end-use for 60 days prior to first exposure to the display devices for all subjects in the experiment. Survey information including household characteristics, existing electricity consumption related behaviors, perceptions about electricity consumption, attitudes etc. collected during the recruiting process. Also monthly kWh data.	Monthly kWh usage for 12 months after the start of the test. Hourly kWh per hour by end-use for all three groups throughout the 12 month testing period. Follow-up in home surveys with study participants in all groups to repeat pre-treatment measurements about household characteristics, electricity consumption related behavior, perceptions about electricity consumption, attitudes, etc. Monthly kWh data.

Protocol 5: Sampling

Protocol 5 poses several questions that must be answered in the process of developing the sample to support the experimental design defined in Protocol 4. The questions in this protocol are intended to guide the development of an appropriate sample design and to lead to a reasonably precise description of the sample design and sampling process. The answers to Protocol 5 for this Category 6 research design are as follows.

1. Are the measurements from the experiment to be extrapolated to the broader utility population?

   No.

   a. If yes, indicate whether the sample will be stratified and what variables will be used in the stratification?

      Sample stratification is not required.

   b. If no, describe the list of customers from which the sampling will be obtained.

      Customers will be recruited from all residential customers within a selected media market.

2. Are precise measurements required for sub-populations of interest?

   No.

   a. If yes, describe the sub-populations for which precise measurements are desired.

      N/A

3. What is the minimum threshold of difference that must be detected by the experiment?

   5% for change in annual electricity consumption

4. What is the acceptable amount of sampling error or statistical precision and acceptable level of statistical confidence (i.e., 90%, 95%, 99%)?

+/- 5% statistical precision with 95% confidence

5. Will customers be randomly assigned to treatment and control conditions or varying levels of factors under study?

   Yes.

   a. If yes, do you expect customers to select themselves into the treatment condition?

      Customers will self-select into the research project, but not into specific treatment or control cells. It is possible for subject households to opt themselves out of the study at any time during the experiment, although incentives will be structured to minimize this.

   b. If so, how will you correct for this selection process in the analysis and sample weighting?

      Characteristics of customers who drop out of the treatment and control groups during the course of the study will be carefully tracked and an effort will be made to control for differences in outmigration using statistical regression.

6. If customers will not be randomly assigned to treatment and control conditions or varying levels of factors under study:

   N/A

   a. Describe the process that will be used to select customers for the treatment group(s).

   b. Describe the process that will be used to select customers for the control group, and explain why this is the best available alternative for creating a non-equivalent control group.

   c. If no control group is used, explain how the change in the outcome variables of interest will be calculated.

7. Describe the sample design that will be used in the study.

   Table 10-3 describes the sample design for this study. All customers in the study will be volunteers. They will be randomly assigned to the treatment and control conditions after they have volunteered.

   As explained in Section 8, the power of statistical tests is greatly magnified when the outcome variable of interest is measured repeatedly (e.g., monthly electricity consumption). Instead of just one measurement of household electricity consumption before the onset of the treatment and one after, there are actually 12 measures of household electricity consumption before and 12 measures of household electricity consumption after the onset of the treatment – one for each month of the study.

   The sample has been designed to estimate the difference in annual electricity consumption for the treatment and control groups to within plus or minus 3% precision with 95% confidence and with 90% statistical power. Practically speaking this design is capable of detecting <u>at least</u> a 3% difference in annual electricity consumption 90% of the time for the following comparisons:

   - *Between households with and without IHDs*
   - *Between households for which the IHD displays information at the premise level and those for which the IHD displays information at the end-use level*

Table 10-3
Sample Design for Energy Impact Analysis

Treatment	Sample Size	Measurements
Enhanced IHD	400	End-use level 15 minute interval measurements of appliances and plug loads.
Enhanced IHD with end-use level display	400	End-use level 15 minute interval measurements of appliances and plug loads.
Control	400	End-use level 15 minute interval measurements of appliances and plug loads.

Protocol 6: Recruitment

As explained above, households for this experimental test will be recruited using a combination of radio, print, and television advertising. Consumers will be assigned to treatment conditions randomly until sufficient samples are assigned to each treatment condition. An additional 10% will be added to each experimental condition to allow for attrition and problems with installation, customer cancellations, and other issues that may arise in carrying out the measurement plan.

1. Is the approach to recruitment for a full-scale program that might ultimately be implemented known with certainty?

   No.

   a. If yes, does the project timeline allow for experimental recruitment to be done in the same manner as the planned recruitment?

   N/A

   b. If yes to Question 1a, what is the recruitment approach that will be used (e.g., direct mail, telemarketing, door-to-door, etc.)?

   N/A

   c. If no to Question 1a, what recruitment options fit within the available timeline?

   In addition to the process that will be deployed in this study, there are three other possible approaches to recruiting subjects: direct mail offers to randomly selected customers, telemarketing to randomly selected customers to find those that are interested, and direct door-to-door "selling" of the study to households that could participate in the study.

   i. What are the potential differences between customers who would be expected to enroll through the long-run recruitment process and customers who would likely enroll through the process that will be used in the experiment?

   The long run recruiting process is unknown.

   ii. Is it possible to recruit a calibration group using the long-run recruitment approach even if they cannot be enrolled in time to be used in the estimation sample for the load impact analysis?

No

2. Is one of the purposes of the experiment to determine what recruitment process works best and, if so, which options will be studied?

   No.

3. Does the sampling plan involve stratification?

   No.

   a. If so, do data exist that allow for stratification prior to recruitment or does the recruitment process need to gather data on customer characteristics and track enrollment according to these criteria?

   N/A

4. What eligibility criteria, if any, apply to each treatment option?

   In order to help maximize the probability that study subjects will remain in the study, eligibility will be limited to homeowners in single family structures who have been located there for at least a year and who state that they do not have plans to move within the next year.

   a. For each treatment option that has eligibility restrictions, do data already exist that allow for precise targeting of eligible customers?

   No, but data can be obtained in conjunction with the recruitment process.

   b. If the answer to Question 4.a is no, does the planned recruitment approach allow for eligibility screening to occur and be tracked as part of the recruitment process?

   Yes.

5. Taking into consideration the cost of each sample point and any other relevant criteria, how important is it to cut off enrollment as close as possible to the target sample size?

   Yes, the premise equipment required to support this study is extensive and requires considerable time and effort to install.

6. If incentives are to be used to enhance subscription, improve persistence, or increase the magnitude of the response to the feedback mechanism, describe the incentives that will be offered and the variations in magnitude of the incentive that will be tested during the experiment.

   Participating customers will be offered a home energy management system and $200 to induce them to participate in the study. There will be no variation in the magnitude of incentives. The incentive is designed to offset the inconvenience associated with the installation of monitoring equipment at their home. Since incentives are not expected to be offered in subsequent marketing, no variations in the impacts of the incentive are being tested.

Protocol 7: Length of Experiment

One objective of the Category 6 feedback experiment is to observe the change in electricity consumption behavior, which varies seasonally. As such, it is important to capture at least one full year of electricity consumption, so that seasonal effects can be observed and the normal variation in seasonal electricity consumption does not lead to erroneous conclusions. In addition, the pre-treatment measurement of electricity consumption by end-use will have been measured in the last two months of the 12 month period preceding the presentation of the feedback information display systems. It is important to compare the behavior of consumers for these two months with their behavior in the last two months of the experimental exposure period to ensure that behaviors were measured during comparable times of the year.

Answers to the questions posed in Protocol 7 are provided below.

1. Is it possible to run the experiment for at least two years?

   Yes, but the experiment will only be run for 12 months since the changes in behavior resulting from exposure to the different treatments is expected to emerge within the first 90 days of treatment and it is impossible to measure impacts on the acquisition behavior or other long term impacts given the sample sizes used in this study.

   a. If no, how will the persistence of the effect be determined?

   The measurement technology used in the study is unobtrusive and if necessary the observational period can be extended beyond 12 months. The decision as to whether the duration of the experiment should be extended can be made during the last two months of the experimental test period.

2. What is the maximum amount of time consumers can be exposed to the feedback mechanism?

   See above.

3. Do pre-treatment data for the relevant variables already exist or must time be allowed to obtain pre-treatment data?

   Yes and No. Pre-treatment data already exist for electricity consumption (i.e., 12 months of monthly electricity consumption information). In addition, pre-treatment survey data will be collected as part of the recruiting process when households are assigned to treatment and control conditions. However, pre-treatment hourly data on electricity consumption by end-use does not exist and will have to be collected during the pre-treatment period.

   a. If pre-treatment data do not already exist, how long must the pre-treatment period be to support the experimental objectives?

   Two months for end-use data.

   b. If pre-treatment data do not already exist, can the experiment be conducted using only post-treatment data, and what adjustments to sample design will be required to employ a post-test-only design?

   N/A

4. What is the expected amount of time required for consumers to receive and understand the information being provided to them?

   Feedback on the rate of energy use and cost is provided in real-time. It is expected that customers will quickly learn the features and functionality of the device. A helpline will be established to assist customers who might have trouble learning how to use the features of the each device.

5. What is the expected amount of time needed by consumers to implement behavioral changes in response to the information provided?

   Changes in usage behavior are likely to occur quickly, but vary seasonally. Changes in purchase behavior, if they exist, occur primarily at the time of the normal turnover in the appliance stock and will not be tracked.

6. How long between the time when a consumer implements a change in behavior and when the feedback associated with that change is likely to be delivered to consumers?

   Feedback on the rate of energy use and cost is in real-time. Changes in cumulative energy use or costs over a period of time are tied to the time period of interest.

7. What is the minimum amount of time the effect of the feedback mechanism must persist to cost-justify investment on the part of the utility?

   To be determined based on the magnitude of savings estimated through the research. Also, it depends on how much costs for the information feedback devices fall in future years and the magnitude of installation costs change with technology. The costs for the equipment used in the study will not be indicative of what long run costs would be.

   a. If the duration of the experiment is shorter than the expected useful life of the measure, how will the determination be made as to whether the effect of the feedback persists long enough to be cost effective?

   The persistence of the effect will be observed over the 12 month interval of the test and will be projected to the useful life of the device.

8. Is the feedback mechanism expected to affect consumers' decisions about the energy efficiency or demand responsiveness of new/replacement appliances?

   The impact of the feedback mechanism on appliance purchase/replacement is unknown and cannot be observed in this study given the duration of the study and sample sizes involved.

   a. If yes, how will the impact of the feedback mechanism on this behavior be measured?

   N/A

9. How much time is needed between when the research plan is completed and approved, and when treatments are in place for experimental participants?

   The primary determinants of the length of time needed before all treatment devices are in place are the time required to select and acquire the feedback devices, hire a recruitment firm and implement the advertising campaign, recruit and screen participants, and schedule and install the devices. This could easily take six months. Once all of this is complete, two months will be required to record end-use data before the treatment devices are allowed to be turned on by customers.

10. How much time is required between when the final data are obtained from the experimental observations and when the analysis can be completed?

    Analysis of the data from the experiment can be completed within one month of the final data collection.

11. What are the drop-dead dates for when draft and final results from the experiment are needed?

Estimates will be developed at various points throughout the duration of the project.

Protocol 8 and 9: Data Requirements and Collection Methods

Protocol 8 contains a table that can be used to summarize the types of data that will be gathered during the experiment, the applicable population, the frequency with which it will be gathered over the research period, the method of data collection or source of the data, any issues that might exist with the data or that might arise as a result of the data collection process, and how those issues will be addressed. Table 10-4 contains the relevant information for the Category 6 feedback options being addressed in this example.

Table 10-4
Protocol 8: Data Requirements and Collection Methods

Electricity consumption	
Description	Monthly kWh, start and end dates for billing period, hourly kWh by end-use, pre-test and post-test observations for all parties in the study.
Population	All treatment and control customers.
Frequency	Monthly kWh for 12 months prior to first treatment through the end of the study period, a total time span of 24 months. Pre-test measurements of electricity consumption by end-use will be collected for the 60 days preceding the delivery of the IHD devices – 1,288 hours. Post-test measurements of electricity consumption by end-use will be collected for 8760 hourly kWh by end-use for all study participants, two survey measurements. Precisely how and how frequently the end-use data will be delivered to the research team will depend on the capabilities of devices and vendors that is yet to be determined.
Method/Source	Utility MDMS/billing system provides monthly kWh for pre-test observations, post-test measurements provided by feedback IHD system provider to be determined.
Issues and solutions	Existing vendors can supply the technology to collect and report both premise level and end-use level data to IHDs. However, care will have to be taken to ensure that functionality of the units is perfectly controlled. Moreover, it may be that household level applications of these technologies that are spread thinly throughout a geographical service territory are incompatible with communications network designs currently contemplated by vendors. It may therefore be necessary to use alternative communications channels that do not rely on back office systems currently offered by vendors. This may entail increased expense to transfer data over other communications systems and development of data base management systems required to support acquisition and management of end-use level data.

Table 10-4 (continued)
Protocol 8: Data Requirements and Collection Methods

Socio-demographic and appliance data	
Description	Customer characteristics (e.g., income, persons per household, size of house), customer perceptions, and reported baseline appliance usage behavior and appliance holdings.
Population	Treatment and control customers.
Frequency	Baseline data will be obtained when subjects are recruited into the experiment. Post-test data will be obtained at the conclusion of the feedback exposure interval (i.e., after 12 months of continuous exposure).
Method/Source	Baseline survey data will be collected during an in-person interview that takes place with a household member when the monitoring systems are installed. The post-test interview will be collected in an in-person interview conducted when the monitoring equipment is removed at the end of the experiment.
Issues and solutions	Response to the baseline survey is mandatory for participation in the study and it will be explained during the recruiting process that a household interview at the conclusion of the study is mandatory. $100 of the $200 incentive for participation in the study will be withheld until after the household interview is concluded.
Energy using behavior	
Description	Data on energy usage behavior (e.g., thermostat settings and habits, number of loads of wash by type (e.g., cold wash, hot wash, etc.), dishwasher usage (number of loads per week, etc.). Same questions asked during the pre-test survey and the post-test survey (that is, customers are not asked to describe how their behavior has changed, just what their behavior is). In addition, hourly electricity consumption measurements by end-use will be collected for 60 days preceding the delivery of the feedback display unit to treatment households and throughout the 12 months period after the feedback display unit is delivered.
Population	Treatment and control customers.
Frequency	Survey data will be collected prior to and immediately after the close of the feedback exposure period. Hourly electricity consumption by end-use will be collected throughout the feedback exposure period.
Method/Source	Same as above.
Issues and solutions	Same as above.

Table 10-4 (continued)
Protocol 8: Data Requirements and Collection Methods

Use of information	
Description	The post-test survey interview will ask treatment subjects whether they are still using the feedback device, what screens they find useful, what information they have ignored, when the last time they consulted the system, whether other household members used it, whether it stimulated them to change anything in the way they operated their home etc.
Population	Treatment customers.
Frequency	Once during the post-test survey.
Method/Source	Same as above.
Issues and solutions	Same as above.
Weather data	
Description	Hourly temperature and humidity for weather stations in close proximity to each customer in control and treatment groups. Will be converted to variables such as cooling and heating degree hours, temperature-humidity index, etc.
Population	All treatment and control customers.
Frequency	Monthly for 12 months prior to first treatment through the end of the study period, a total time span of 24 months.
Method/Source	NOAA and/or other public weather data sources.
Issues and solutions	Careful attention must be paid to geography and micro-climates when assigning customers to weather stations.
Other	
Description	Additional information available from the utility that could be used as explanatory variables in regression models that determine the change in electricity consumption, or to identify high responder customers, and/or to detect non-response bias in surveys, etc., would include such things as prior participation in utility sponsored EE and DR programs, tariff, location (for mapping with weather stations and perhaps with publicly available data such as census data), etc.
Population	All treatment and control customers.
Frequency	Updated on a regular basis (perhaps quarterly) throughout the study period.
Method/Source	Varies – see “Description” section above.
Issues and solutions	None.

Protocol 9 lists a common set of data that EPRI recommends be gathered for each experiment and asks research planners to identify which of the recommended minimum requirements will NOT be included as part of the data collection efforts associated with an experiment.

In order to enhance cross-utility comparisons of experimental results or to allow for data pooling across experiments, the following data should be obtained for each experimental subject. Please indicate if any of the data elements are NOT going to be obtained.

> *ALL OF THE DATA LISTED BELOW WILL BE OBTAINED FOR TREATMENT AND CONTROL CUSTOMERS.*

1. Designator indicating the treatment to which the observation was assigned (e.g., Treatment 1, Treatment 2, Control, etc.)
2. For customers in all experiments that do not involve interval metering:
   a. kWh usage for all pre-treatment and treatment billing periods for each participant
   b. Meter read date for each billing period
   c. Monthly electricity bill
   d. Tariff designation
   e. Date that treatment went into effect for all treatment customers
   f. Date customer left experiment for each customer that left before the end of the treatment period
3. For customers in all experiments involving demand-metered customers, in addition to all of the data in Question 1 above:
   a. Monthly peak demand
4. For customers in all experiments in which all customers have interval meters
   a. kWh usage for each hour for the pre-treatment and treatment time periods
   b. Items 1b, 1c, 1d, and 1e
5. For customers in all experiments, data on the following customer characteristics:
   a. Zip code
   b. Date the customer entered the experiment (treatments or controls)
   c. Date the customer departed from the experiment (treatments or controls)
   d. Reason the customer departed from the experiment (treatments or controls)
   e. Presence of central air conditioning
   f. Number of room air conditioners
   g. Presence of electric space heating by type (e.g., base board, heat pumps, etc.)
   h. Type of control device for air conditioning and space heating (e.g., standard thermostat, programmable thermostat, etc.)
   i. Presence of electric water heating by type (e.g., tank, tankless, etc.)
   j. Presence of dishwasher, clothes washer, electric drier, electric cook top, electric oven, electric hot tub/Jacuzzi, swimming pool pump, domestic water pump, Plasma TV
   k. Housing type (e.g., single family detached, single family attached, multi-family, etc.)
   l. Size of dwelling

m. Number of persons in household by age grouping

n. Annual household income

Protocol 10: Key Support Systems

Another key element of research design is determining the key systems and materials that will be needed to support an experiment and how those needs will be fulfilled. Protocol 10 contains a table that can be used to identify the key systems and materials that will be needed, delineate the primary fulfillment plan for each, identify any risks that exist and, if relevant, a backup plan. Given the nature of the Category 6 example, most of the needed support can be outsourced and there are few significant risks with fulfillment. Table 10-5 shows how Protocol 10 would be completed for the Category 6 example presented here.

Table 10-5
Key Support Systems and Materials Inventory and Assessment

Description	Fulfillment Plan	Summary of Risks	Alternative Options
Metering	Standard watt hour meters are sufficient to measure change in electricity consumption. However, an ancillary data collection system will have to be installed to measure electricity consumption by end-use.	Both the IHD and communications technologies required to support the experiment are new and not tested on a large scale. Hawthorne may be an issue, particularly with participants who have end-uses being monitored during the 60 day pre-period. Also, would need to “activate” feedback/IHD after 60-day period.	A small scale hardware test should be carried out with at least ten installations prior to implementation of the full scale experiment. Surveys administered at key time intervals to try to assess Hawthorne? May require having installer come to home for second visit.
Meter Data Management	Standard data management for monthly consumption is adequate.	It may be necessary to develop or procure a data base management system to support measurements of energy consumption by end-use.	N/A
Billing	Standard	None	N/A
Information Treatments	To ensure comparability of the experimental stimulus, a single IHD display unit will be used. The capability to display energy use by end-use will be disabled for the IHD enhanced condition.	Control group customers may acquire measurement capabilities for displaying household loads during the experiment.	An effort must be made during the exit interviews to observe whether the household has procured the capability to monitor their loads during the course of the experiment.

Table 10-5 (continued)
Key Support Systems and Materials Inventory and Assessment

Description	Fulfillment Plan	Summary of Risks	Alternative Options
Recruitment Tracking	It will be necessary to develop data collection forms, a data base management system for tracking in-take survey responses, equipment serial numbers assigned to addresses, data channel ID numbers assigned to end-uses, installation dates, interview dates and equipment recover dates, and other identifying information used in the recruiting process.	Must ensure that data is captured and communicated internally and to outsourcing and evaluation contractors.	N/A
Recruitment Process	Employ a research firm that specializes in recruiting subjects for medical clinical trials to recruit subjects.	Most firms involved in clinical trials have no experience with recruiting for this kind of study. This may lead to over or under marketing problems.	Do a small scale test in a remote community to gauge community reaction to marketing efforts.
Marketing Material	No marketing material is needed beyond the advertising required to recruit participants.	N/A	N/A
Customer Information/ Education Materials	No information or education materials beyond those that are contained in the descriptive materials provided by IHD providers.	N/A	N/A
Customer Support	A toll free number will be provided to treatment customers so that they can call if they have questions about how to use the device.	Some risk that device suppliers won't meet the service standards that would be ideal from a utility's perspective. A decision will be made as part of the device procurement process concerning whether the device suppliers can or should be used for this or whether a utility technical expert would be better.	Utility staff could be used to man the technical hotline.
Surveys	Pre-treatment and post-treatment surveys must be developed.	The response to post-treatment surveys may be lower than assumed. A 10% contingency has been added to all sample cells to account for outmigration from the study.	Increase contingency sample size to 20%.
Other	N/A	N/A	N/A

Protocol 11: Analysis Plan

The research described in this Category 6 example has several objectives:

1. Measure the change in annual household energy use that occurs when households are provided with an enhanced IHD and an IHD that provides real-time end-use usage data, and the difference in impacts between these options.
2. Identify the differences in behavior that occur as a result of the installation of the IHD and IHD with display at an end-use level.
3. Identify the ways in which consumers use both kinds of devices. In particular:
    a. What functionality do they use/ignore?
    b. Do they use the display devices throughout the measurement period?
    c. If they don't, about how long do they use them before they cease to pay attention to the information they are providing?

Different analysis methods will be used to answer the above questions. To estimate the effects of providing IHDs to customers, panel regression models will be used, including coefficients for factors that are changing over time (principally duration of exposure to the IHD and weather). In the model, there are 12 pre-test measurements and 12 post-test measurements for the groups and for the control group.

A similar approach will be used to identify differences in electricity consumption before and after exposure to the IHD. However, in this case, the electricity consumption for the households will be disaggregated into end-uses. That is, the above analysis will be carried out for each major household end-use. There are approximately 650 hourly measurements before the IHDs are installed and 650 measurements taken approximately ten months after the IHDs have been installed (exactly 12 months after the first set of measurements). It makes sense to aggregate the hourly intervals into meaningful intervals for assessing the impacts of the treatment conditions on energy consumption per end-use. There are three possibilities – daily, weekly, and monthly. The analysis will be conducted for all three levels of aggregation. That is, impacts on daily, weekly, and monthly energy consumption per end-use will be identified.

The survey data will also provide information that can be used to measure the impacts of the treatments. Household energy use related behavior will be measured at two intervals. The extent of change in each of the indicators of household energy use behavior will be assessed using a difference of differences calculation.

The analysis of behavioral change will also involve a two-stage modeling approach. In the first stage, changes in behavioral variables between the first and second surveys for each individual will be calculated. These differences will then be used as dependent variables in second stage regressions that relate customer characteristics to changes in behavior. In this manner, one might find, for example, that households with the most significant changes in thermostat settings are also households that participate more in EE programs, or households that have two working members and no children.

Budget

The cost estimates described below are not necessarily indicative of the current market prices of the equipment and services that would be required to actually carry out the study described in this section. They are meant as placeholders describing the categories of costs that must be considered and the level of detail required for planning.

Design consultant	$50,000 to $75,000
Recruitment	$540,000 ($200 per recruit + $300 interview and install premise equipment x 1,200 installations)
Premise Equipment	$1,800,000 ($1500 per installation x 1,200)
Surveys	$360,000 (1200 x $300)
Analysis	$200,000 to $300,000
Total cost	$2.95 m to $3.08m over roughly 30 months

The costs above may be more than many utilities would be able to spend. This is typical of research planning, where the preferred design must be reconfigured once budget realities are revealed. There are various ways of reducing the costs, including testing fewer treatments (eliminating the 12 month test, Treatment 1b, for example), less precision (thus reducing required sample sizes and survey costs), fewer panels (which would compromise the ability to detect seasonal effects or persistence), and others.

Schedule

Overall, the project schedule will require at least six months from the time of project approval until when treatments are in place, 12 months for the treatment period, and one to two months to complete the analysis and produce a report.

A
APPENDIX: BLANK PROTOCOL TABLES AND SHEETS

Protocol 1: Defining Information Feedback Treatments

Please complete the following table. When describing the information content that will be made available for each treatment, include a detailed description for Treatment 1 and then define differences in the content between Treatment 1 and the other treatments, rather than repeating the same portions of the description when content overlaps across treatment options. If more than three treatment/segment combinations are to be tested, additional tables should be completed until all treatment/segment combinations are identified.

ATTRIBUTE	TREATMENT 1	TREATMENT 2	TREATMENT 3
INFORMATION CONTENT			
Delineate all content for Treatment 1	Detailed description	State the content that is different from Treatment 1	State the content that is different from Treatment 1
INFORMATION FORMAT			
Numerical (toggle through each output)	Y/N?	Y/N?	Y/N?
Tabular	Y/N?	Y/N?	Y/N?
Graphical	Y/N?	Y/N?	Y/N?
Other	Describe	Describe	Describe
DELIVERY CHANNEL			
Dedicated IHD, Professionally Installed	Y/N?	Y/N?	Y/N?
Dedicated IHD, Customer Installed	Y/N?	Y/N?	Y/N?
PCT	Y/N?	Y/N?	Y/N?
Pushed to PC/TV through USB Device	Y/N?	Y/N?	Y/N?
Customer Access through Web Portal	Y/N?	Y/N?	Y/N?
Other	Describe	Describe	Describe
INTERACTIVE FEATURES			
Describe in detail any interactive features provided for each treatment	Detailed description	State the content that is different from Treatment 1	State the content that is different from Treatment 1
DELIVERY FREQUENCY			
Frequency	Describe	Describe	Describe

Protocol 2: Determining Outcome Variables to be Measured

Please provide answers to the following questions as part of the planning process.

1. Which of the following outcome variables will the experiment be designed to measure? If the outcomes of interest vary by customer segment, indicate the desired outcomes for each customer segment delineated in question 1.
   a. Change in annual kWh
   b. Change in monthly kWh (designate whether for each month or for selected months)
   c. Change in hourly or sub-hourly kWh (designate sub-hourly intervals) for each hour (or sub-hour) for specific, designated time periods, (delineate time periods, e.g., all hours in the year, all hours in selected months, all hours on selected days within a month such as system peak days, etc.).
   d. Change in peak demand (kW) for specific, designated times (delineate times, e.g., at time of annual system peak, for each monthly system peak, etc.).
2. Will the experiment seek to identify and quantify the prevalence of the specific types of behavior that change as a result of the treatment? If yes, delineate whether any specific types of behavior are of particular interest (e.g., increase thermostat set point in summer, turn off lights more, etc.).
3. Will the experiment seek to understand how consumers process and use the information being provided to change their behavior?
4. Will the experiment seek to understand the key drivers of customer choice associated with various information options and program/marketing methods? If yes, describe the various marketing strategies/offers that will be tested for each information option and market segment.

Protocol 3: Delineating Customer Sub-Segments of Interest

Please complete the following table, indicating the population sub-segments of interest and the a priori assumption concerning how outcomes for each segment might differ from other segments of interest.

Customer Sub-Segment Description	Hypothesis
Example: Low income consumers	Low income consumers have less discretionary loads and, therefore, are expected to have lower percentage and absolute reductions in annual energy use
(Describe)	(State Hypotheses)
(Describe)	(State Hypotheses)
(Describe)	(State Hypotheses)
(Add additional rows as needed)	

Protocol 4: Defining the Experimental Design

Please provide answers to the following questions as input to experimental design.

1. Will pre-treatment data be used?
2. Do the appropriate data already exist on all relevant customers, or do meters or other equipment need to be installed in order to gather pre-treatment data?
3. How long of a pre-treatment period of data collection is required?
4. Is a control group (or groups) required for the experiment?[64]
5. Is it possible to randomly assign observations to treatment and control groups?
6. If random assignment is either inappropriate (e.g., if customers are expected to self-select into the program in the future) or impossible to achieve, how will a suitable control group be selected?
7. Using the framework outlined in Section 3 describe treatment(s) and blocks (if any) that will be used during the feedback experiment. This description should be a variation on Figure 3-2 which shows an example of how treatments (and control groups) will be measured for a simple experiment involving two treatments, a control group, and two sampling strata.

Protocol 5: Defining the Sampling Plan

Please answer the following questions pertaining to sample planning.

1. Are the measurements from the experiment to be extrapolated to the broader utility population?

[64] This will almost always be the case, but there are circumstances where other quasi-experimental design techniques can be safely substituted for a control group. See Sullivan, 2009.

   a. If yes, indicate whether the sample will be stratified and what variables will be used in the stratification.

   b. If no, describe the list of customers from which the sampling will be obtained.

2. Are precise measurements required for sub-populations of interest?

   a. If yes, describe the sub-populations for which precise measurements are desired.

3. What is the minimum threshold of difference that must be detected by the experiment?

4. What is the acceptable amount of sampling error or statistical precision and acceptable level of statistical confidence (i.e., 90%, 95%, 99%)?

5. Will customers be randomly assigned to treatment and control conditions or varying levels of factors under study?

   a. If yes, do you expect customers to select themselves into the treatment condition?

   b. If so, how will you correct for this selection process in the analysis and sample weighting?

6. If customers will not be randomly assigned to treatment and control conditions or varying levels of factors under study:

   a. Describe the process that will be used to select customers for the treatment group(s).

   b. Describe the process that will be used to select customers for the control group, and explain why this is the best available alternative for creating a non-equivalent control group.

   c. If no control group is used, explain how the change in the outcome variables of interest will be calculated.

Protocol 6: Identifying the Recruitment Strategy

Please answer the following questions pertaining to recruitment.

1. Is the approach to recruitment for a full-scale program that might ultimately be implemented known with certainty?

   a. If yes, does the project timeline allow for experimental recruitment to be done in the same manner as the planned recruitment?

   b. If yes to Question 1a, what is the recruitment approach that will be used (e.g., direct mail, telemarketing, door-to-door, etc.)?

   c. If no to Question 1a, what recruitment options fit within the available timeline?

      i. What are the potential differences between customers who would be expected to enroll through the long-run recruitment process and customers who would likely enroll through the process that will be used in the experiment?

   ii. Is it possible to recruit a calibration group using the long-run recruitment approach even if they cannot be enrolled in time to be used in the estimation sample for the load impact analysis?[65]

2. Is one of the purposes of the experiment to determine what recruitment process works best and, if so, which options will be studied?
3. Does the sampling plan involve stratification?
   a. If so, do data exist that allow for stratification prior to recruitment or does the recruitment process need to gather data on customer characteristics and track enrollment according to these criteria?
4. What eligibility criteria, if any, apply to each treatment option?
   a. For each treatment option that has eligibility restrictions, do data already exist that allow for precise targeting of eligible customers?
   b. If the answer to Question 4a is no, does the planned recruitment approach allow for eligibility screening to occur and be tracked as part of the recruitment process?
5. Taking into consideration the cost of each sample point and any other relevant criteria, how important is it to cut off enrollment as close as possible to the target sample size?
6. If incentives are to be used to enhance subscription, improve persistence, or increase the magnitude of the response to the feedback mechanism, describe the incentives that will be offered and the variations in magnitude of the incentive that will be tested during the experiment.

Protocol 7: Identifying the Length of the Experiment

Please answer the following questions pertaining to the experimental time frame.

1. Is it possible to run the experiment for at least two years?
   a. If no, how will the persistence of the effect be determined?
2. What is the maximum amount of time consumers can be exposed to the feedback mechanism?
3. Do pre-treatment data for the relevant variables already exist or must time be allowed to obtain pre-treatment data?
   a. If pre-treatment data do not already exist, how long must the pre-treatment period be to support the experimental objectives?
   b. If pre-treatment data do not already exist, can the experiment be conducted using only post-treatment data, and what adjustments to sample design will be required to employ a post-test-only design?[66]

[65] For example, it might be necessary to recruit by telephone in order to meet a deadline to install meters prior to a summer season when treatments must go into effect. However, in parallel with this effort, it could be useful to recruit a small sample of customers using direct mail, even though it would not be possible to enroll them and install meters prior to the start of the treatment period. The characteristics of this small calibration group could then be compared with those of the group recruited through telemarketing to determine whether there are observable differences in the two groups that might affect the impact estimates obtained from the telemarketing recruitment process.

4. What is the expected amount of time required for consumers to receive and understand the information being provided to them?[67]
5. What is the expected amount of time needed by consumers to implement behavioral changes in response to the information provided?
6. How long between the time when a consumer implements a change in behavior and when the feedback associated with that change is likely to be delivered to consumers?[68]
7. What is the minimum amount of time the effect of the feedback mechanism must persist to cost-justify investment on the part of the utility?

    a. If the duration of the experiment is shorter than the expected useful life of the measure, how will the determination be made as to whether the effect of the feedback persists long enough to be cost effective?

8. Is the feedback mechanism expected to affect consumers' decisions about the energy efficiency or demand responsiveness of new/replacement appliances?

    b. If yes, how will the impact of the feedback mechanism on this behavior be measured?

9. How much time is needed between when the research plan is completed and approved, and when treatments are in place for experimental participants?
10. How much time is required between when the final data are obtained from the experimental observations and when the analysis can be completed?
11. What are the drop-dead dates for when draft and final results from the experiment are needed?

Protocol 8: Identifying Data Requirements and Collection Methods

Energy Use	
Description	
Population	
Frequency	
Method/Source	
Issues and solutions	
Socio-demographic and appliance data	
Description	

[66] Put another way, are pre-treatment data essential or is there a "work around" that can be used if the experimental time frame does not allow for the collection of pre-treatment data?

[67] For real-time feedback, this time period is likely to be measured in days. For monthly information provision, it could take several months before consumers would receive sufficient information feedback to factor it into their usage decisions.

[68] With real-time feedback, the time required for consumers to observe the impact of a change in behavior is almost instantaneous whereas for monthly feedback, it may take several months to see the affect of a change.

Population	
Frequency	
Method/Source	
Issues and solutions	
Energy using behavior	
Description	
Population	
Frequency	
Method/Source	
Issues and solutions	
Use of information	
Description	
Population	
Frequency	
Method/Source	
Issues and solutions	
Weather data	
Description	
Population	
Frequency	
Method/Source	
Issues and solutions	
Other	
Description	
Population	
Frequency	
Method/Source	
Issues and solutions	

Protocol 9: Meeting Minimum Data Requirements for Cross-Utility Comparisons and Pooling

In order to enhance cross-utility comparisons of experimental results or to allow for data pooling across experiments, the following data should be obtained for each experimental subject.

1. Designator indicating the treatment to which the observation was assigned (e.g., Treatment 1, Treatment 2, Control, etc.)
2. For customers in all experiments that do not involve interval metering:
   a. kWh usage for all pre-treatment and treatment billing periods for each participant
   b. Meter read date for each billing period
   c. Monthly electricity bill
   d. Tariff designation
   e. Date that treatment went into effect for each treatment customer
   f. Date customer left experiment for each customer that left before the end of the treatment period
3. For customers in all experiments involving demand-metered customers, in addition to all of the data in Question 1 above:
   a. Monthly peak demand
4. For customers in all experiments in which all customers have interval meters:
   a. kWh usage for each hour for the pre-treatment and treatment time periods
   b. Items 1b, 1c, 1d, and 1e
5. For customers in all experiments, data on the following customer characteristics:

Variable	Specification
Zip code	5 digit
Date customer entered the experiment	mm/dd/yy
Date customer departed the experiment	mm/dd//yy
Reason customer withdrew from experiment	Text (e.g., deceased, moved, etc.)
Air conditioning systems	Number of central AC units Number of room AC units
Space heating systems	Presence of electric baseboards (Y/N) Number of central heating systems (gas) Number of central heating systems (electric)
Type of space heating system control	Manual Standard thermostat Programmable thermostat
Water heating systems	Electric Gas Solar
Household appliance inventory	Number of the following appliances:

Variable	Specification
	Home computers Printers Dishwashers Clothes washers Electric dryers Electric cook tops Electric ovens Electric spas Pool pumps Domestic water pumps CRT TVs Plasma TVs LED TVs
Dwelling type	Single family detached Single family attached (e.g., duplex or town house) Multifamily (e.g., apartment or condo) Manufactured home (e.g. mobile home) Other
Dwelling size	Sq. ft of enclosed area
Number of persons in household by age group	Age 1-6 7-19 20-24 25-60 61-70 > 70
Annual household income	For the year preceding the start of the experiment

Protocol 10: Identifying Key Support Systems and Materials

Please complete the following table. Enter N/A (not applicable) for systems and materials that are not needed for the experiment being designed.

Description	Fulfillment Plan	Summary of Risks	Alternative Options
Metering			
Meter Data Management			
Billing			
Information Treatments			
Recruitment Tracking			
Recruitment Process			
Marketing Material			
Customer Information/ Education Materials			
Customer Support			
Surveys			
Other			

Protocol 11: Load Impact Analysis

For analyses based on the difference-in-differences approach using pre- and post-measurements for treatment and control groups, produce the following information for the average customer for each treatment tested:

1. The mean and standard deviation for the treatment and control group for each strata or customer segment delineated in the experiment, and for the group as a whole, for each time period (e.g., annual kWh, monthly kWh, average weekday kWh, peak hour for each monthly system peak day, etc.)[69] for the pre-treatment and treatment time periods
2. The number of customers included in each calculation in Question 1
3. The estimated impact and the standard error of the estimated impact for each period, for each strata or customer segment delineated in the experiment, and for the group as a whole and the value of the appropriate measure of statistical significance of the impact (e.g., the t-statistic)
4. For experiments involving stratification of customers, estimate the difference in load impacts across strata and the value of the appropriate measure of statistical significance of any difference across group

[69] Also report how each relevant period is defined. For example, for experiments involving kWh meters, how is a month defined in light of the fact that nearly all billing cycles straddle calendar months?

5. For each time period for which a load impact is reported, estimate the cooling degree hours to base 72°F and the heating degree hours to base 65°F
6. Calculate the average values and standard deviations for all customer characteristics data gathered for each treatment and control group used in the calculations
7. Calculate whether there are statistically significant differences in all characteristics for which data are gathered between treatment and control groups, and between customers in each stratum

For analyses involving repeated measures or regression modeling, produce the following:

8. Definitions for all variables used in all estimated regressions, a description of the functional form of the equations, and an explanation of logic underlying inclusion of all variables[70]
9. A print out of all regression results showing the estimated coefficients, r-squared values, and other relevant statistics provided through standard statistical software packages
10. The estimated impact and the standard error of the estimated impact for each period, for each strata or customer segment delineated in the experiment, and for the group as a whole and the value of the appropriate measure of statistical significance of the impact (e.g., the t-statistic)
11. The estimated value of load impacts based on long-term normal weather conditions, and the definition of how long-term normal weather is defined[71]
12. For experiments involving stratification of customers, estimate the difference in load impacts across strata and the value of the appropriate measure of statistical significance of any difference across groups

Protocol 12: Behavioral Change Analysis

1. Are estimates of the rates of adoption of feedback technology or program (the treatments) required as part of the research? If yes:
    a. Describe the data that will be collected to measure the rate of acceptance for each treatment (including any data that must be acquired from third party vendors or surveying).
    b. Describe the statistical techniques that will be used to describe the impacts of customer characteristics (e.g. household lifestyle) and feedback system characteristics (e.g., price) on rate of acceptance.
2. Will end-use metering be used to describe changes in electricity consumption behavior by end-use? If yes:
    a. List the end-uses that will be metered for treatment and control households (e.g. lighting, heating, ventilating, and air conditioning (HVAC), dish washing, etc.).

[70] For example,

month I	Dummy variables for month of the year, designed to pick up seasonal effects
dayofweeki	Dummy variables designed to pick up day-of-week effects

[71] This requirement assumes that weather terms are properly included in the regression models, in which case producing estimates is quite straightforward. Weather normalization is not indicated for non-regression based calculations as providing weather normalized estimates is not a trivial extension of the estimation method.

   b. For each end-use, indicate whether there is an a priori hypothesis concerning how the feedback mechanism may affect the end-use.

   c. Describe the technology that will be used to record and recover end-use measurements.

   d. Generally describe the analysis technique that will be used to identify changes in energy consumption by end-use.

3. Will statistical surveys be used to measure changes in electricity consumption related behavior or appliance acquisition behavior? If yes:

   a. Describe how the surveys will be implemented, including:

      i. Whether the surveys will consist of panels (i.e., repeated measurements of the same subject) or cross-sections or both

      ii. How often before and during the test the customers will be contacted

      iii. Measures that are taken in survey design to measure and control for selection effects (due to survey non-response and Hawthorne effects)

   b. Generally describe the analysis techniques that will be used to identify changes in consumer perceptions and behavior using the survey data.

Protocol 13: Analysis of Participant Use of Information Feedback

This protocol is only to be completed for studies that are intended to analyze the ways in which consumers use the information that is provided in feedback.

1. Will customers in treatment group(s) be surveyed to study the ways in which consumers used the feedback? If no:

   a. Describe the procedures that will be used to measure the ways in which consumers are using the information provided by the feedback mechanism.

2. If more than one treatment is under study, will a common survey be used in all treatments? If no:

   a. How will consumer responses from the different treatments be compared?

3. Will this survey be carried out during the time that the treatment is taking place? If yes:

   a. What actions will be taken to ensure that this survey does not cause a Hawthorne effect when measuring the effect of the treatment on electricity consumption?

4. How will consumers be selected for the survey?

5. List the research questions the survey is intended to address (e.g. “what screens do consumers find most useful?”).

6. List the survey questions that will be asked of consumers to address the research questions (e.g., “Thinking of the times when you have used the (insert feedback device name), what screen do you think provides you with the most useful information?”).

7. Describe the statistics that will be used to summarize the responses of customers.

Protocol 14: Documentation of Feedback Experiments

This protocol is intended to standardize the reporting of certain critical information that is needed to understand and interpret the results of a feedback experiment. Reports concerning feedback experiments should contain the following information.

1. An executive summary including:
    a. A description of the study objectives
    b. An overview of the experimental design
    c. A description of the rate of acceptance of the feedback mechanism during test marketing
    d. A description of the impacts of the feedback mechanism on energy consumption and/or demand in percentage terms along with an indication of the calculated upper and lower confidence levels associated with reported point estimates
    e. An executive summary that clearly describes any changes that were observed, if an effort was made to observe changes in consumer behavior or device usage patterns
2. A description of the feedback mechanisms that were tested along with an explanation of how these mechanisms are supposed to alter consumer behavior – This description should clearly describe the functionality of any equipment that was tested as well as a description of any other experimental factors that were included in the test such as variations, incentives, or other information provided to customers during the tests.
3. A detailed description of the experimental design that was used in the study, including:
    a. All variations in marketing strategies tested or used
    b. Delivery mechanism/hardware combinations tested
    c. Any variations in incentives used to enhance recruitment, persistence, and performance
    d. Variations in other factors (e.g., supplemental information or training) that may have been tested during the study
4. A detailed description of the sample design, and sampling process, including:
    a. The population of interest
    b. The sampling frame
    c. Stratification design if any
    d. Allocation of initial sample to treatment and control conditions
    e. Allocation of final recruited sample to treatment and control conditions
    f. Analysis of selection bias that may have occurred during the sampling process, if any
5. A description of the historical timeline of the test, including:
    a. Planning phase
    b. Operational phase
6. A detailed discussion of the statistical procedures used in the analysis of the data from the study, including:

   a. Detailed specifications of statistical models used to describe experimental outcomes
   b. Data cleaning procedures used in the study
   c. Procedures used to control for censoring
   d. Procedures used to control for selection (if appropriate)
   e. Weighting procedures used and sampling weights

7. Results reported according to the requirements of the analysis Protocols 11-13.

The Electric Power Research Institute Inc., (EPRI, www.epri.com) conducts research and development relating to the generation, delivery and use of electricity for the benefit of the public. An independent, nonprofit organization, EPRI brings together its scientists and engineers as well as experts from academia and industry to help address challenges in electricity, including reliability, efficiency, health, safety and the environment. EPRI also provides technology, policy and economic analyses to drive long-range research and development planning, and supports research in emerging technologies. EPRI's members represent more than 90 percent of the electricity generated and delivered in the United States, and international participation extends to 40 countries. EPRI's principal offices and laboratories are located in Palo Alto, Calif.; Charlotte, N.C.; Knoxville, Tenn.; and Lenox, Mass.

Together...Shaping the Future of Electricity

Program:

End-Use Energy Efficiency

1020855

Electric Power Research Institute
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com