EPRI | ELECTRIC POWER RESEARCH INSTITUTE

Designing Survey Instruments for Measuring Behavior Change: Applications to Feedback Research

1021961

Designing Survey Instruments for Measuring Behavior Change: Applications to Feedback Research

1021961

Technical Update, December 2011

EPRI Project Manager

J. 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

THE FOLLOWING ORGANIZATION, UNDER CONTRACT TO EPRI, PREPARED THIS REPORT:

Opinion Dynamics Corporation

This is an EPRI Technical Update report. A Technical Update report is intended as an informal report of continuing research, a meeting, or a topical study. It is not a final EPRI technical report.

NOTE

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

ACKNOWLEDGMENTS

The following organization, under contract to the Electric Power Research Institute (EPRI), prepared this report:

Opinion Dynamics Corporation
1999 Harrison Street, Suite 1420
Oakland, CA 94612

Principal Investigators
A. Dougherty
J. Mitchell-Jackson
T. Buhr
K. Randazzo

This report describes research sponsored by EPRI.

This publication is a corporate document that should be cited in the literature in the following manner:

Designing Survey Instruments for Measuring Behavior Change: Applications to Feedback Research. EPRI, Palo Alto, CA: 2011. 1021961.

ABSTRACT

This report provides guidance on how to develop survey-based studies and instruments to gather self-reported changes in participant behavior that can be attributable to feedback interventions. The report provides guidelines, intended to complement past EPRI work on behavioral research evaluation, to move researchers through the study design process, from overarching considerations, to setting up the survey research methods to establish attribution, to providing techniques for collecting behavioral information from a survey instrument. Specifically, this report includes key definitions used when discussing survey-based research as well as key questions that should be asked when considering survey design, data collection methods, and data analysis processes. In addition, this report describes the advantages and disadvantages of using different types of survey study designs and data collection methods. Specific tactics for collecting information on behaviors and minimizing bias through a well-designed survey instrument are also described in this report. Finally, the report provides examples of survey questions as well as a survey review checklist for helping researchers draft an appropriate survey instrument for the purpose of collecting data on behavior and behavior changes. Although this report focuses on the specific application of feedback, it is important to note that the report is also relevant to other types of behavioral programs.

Keywords
Behavior change
Behavioral programs
Feedback
Self-report
Surveys

CONTENTS

1
INTRODUCTION

Study Background

Feedback research evaluations have reported average savings ranging from 0% to 7%[1] depending on the type of feedback intervention; however, questions remain as to which behaviors exactly contribute to these observed savings. Not all savings are the same. There are important distinctions between conservation behaviors and efficiency installations when considering the persistence of savings, when resource planning, and when developing long-term strategies for these programs.

Currently, some data collection techniques, such as appliance-level metering or the use of lighting loggers, can determine exactly how participant daily electricity usage changes in response to feedback interventions. However, these measurement techniques can be expensive for utilities to implement[2] and potentially invasive for participants. Purchase behavior is easier to track, but only if taken through official utility incentive programs or through some other utility program effort.

Given the difficulty of measuring behavior change, survey research remains an important data collection method for quantifying the type, level, and quality of behavior changes associated with feedback interventions, despite the well-understood shortcomings of self-report in our industry.[3] For this reason, it is important to examine how survey instruments should be developed to help researchers design surveys that accurately collect information on behavior changes attributable to feedback interventions.

This report provides guidance on how to develop survey-based studies and instruments to gather self-reported changes in participant behavior that may be attributable to feedback interventions. Here, we provide guidelines designed to move researchers through the study design process, from over-arching considerations, to setting up the survey research methods to establish attribution, to providing specific tactics and techniques for collecting behavioral information

---

[1] EPRI. (2012, Forthcoming). *Understanding Electric Utility Customers: What We Know and What We Need to Know (DRAFT).* Palo Alto, CA: EPRI. Report No. 1023562.

[2] This may change as new innovations in customer-oriented end-use monitoring and control technology evolve.

[3] Ridge, R., Willems, P., Itron, J. F., & Randazzo, K. V. (2009). "The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating the Net-to-Gross Ratio". *International Energy Program Evaluation Conference*. Portland.

Megdal, L. M., Pertusiello, S., & Jacobson, B. (1996). *All These Years Measuring Free Ridership and Now We Measure a Portion of These as Caused by Market Transformation.* Arizona: Presented to the Association of Energy Services Professionals.

Megdal, L. M., Pertusiello, S., & Jacobson, B. (1997). *Measuring Market Transformation Due to Prior Utility Efforts.* Energy Evaluation Conference. Chicago.

Lutzenhiser, L. (2002). *An Exploratory Analysis of Residential Electricity Conservation Survey and Billing Data.* Pullman, WA: Presented to the California Public Utilities Commission.

Hall, N., De Cotis, P., Kushler, M., Megdal, L., & Vine, E. (2007). *An Evaluation Approach for Assessing Program Performance from the State Energy Program.* Washington, DC: Presented to the Office of Weatherization and Intergovernmental Programs within the U.S. Department of Energy.

from a survey instrument. These guidelines are intended to complement past EPRI work on behavioral research design guidelines.[4] As such, this work revisits concepts related to research design and methods, but only as they relate to considerations for designing and fielding survey research. While this report focuses on the specific application of feedback, it is important to note that this report is also relevant to other types of behavioral programs—that is, any program designed to change customer behavior as a result of an intervention.

Please note that this report does not make an argument for using surveys to attribute energy savings to feedback programs; rather, it offers guidelines for those who are interested in measuring behaviors attributable to feedback (and to the extent desired, savings) or for those who just want to pay close attention to assessing behaviors via surveys. The specific applications of the findings gathered through survey research, and the limitations of those findings, should be determined by the researchers within the existing frameworks and protocols of their institutions, jurisdictions, and regulatory environments.

Why Guidelines are Needed

While there is a wealth of literature identifying guidelines for survey design in general, most do not focus on surveys specifically designed to measure behavior and behavior change. Also, there is little research related to the development of surveys intended to assess electricity usage behavior change in particular—a unique challenge given many usage behaviors are often routine and therefore difficult to recall.

Currently, survey research is one of the most commonly used tools employed by researchers in electric energy; it is ubiquitous in market research as well as in evaluations. Unfortunately, "despite the prominent role that surveys play in energy evaluation, survey design is not often given the same level of attention as other needed skills."[5] Without proper structure, surveys can provide misleading, biased, or altogether incorrect information regarding electric energy usage behaviors.

We begin in this report by setting up key definitions for feedback, behavior, and surveys to scope these guidelines and provide readers with a clear understanding of how these terms will be used throughout. In Section 2 we provide key questions that should be asked to scope a behavioral survey effort. These questions should be considered before and during the survey design, data collection, and analysis process. In Section 3, we explain the benefits and drawbacks of different survey study designs, from pre-post and comparison group data collection methods to sampling considerations. We explain the varying levels of rigor associated with each approach and the trade-offs made when choosing one method over another. In Section 4, we move on to discuss specific tactics for collecting information on behaviors and changes in behaviors through survey questions, including a discussion of self-reported attribution approaches. In Section 5, we discuss techniques for reducing bias in question wording and instrument development. We then summarize our guidelines in Section 6 and provide an example instrument in Appendix B and a survey review checklist in Appendix C.

---

[4] EPRI. (2010). *Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols*. Palo Alto, CA: EPRI. Report No. 1020855.

[5] Buhr, T. (2011). "Think Before You Do: The Importance of Survey Design in Program Evaluation". *International Energy Program Evaluation Conference.* Boston, MA.

Key Definitions

For the sake of clarity and scope, we offer the following definitions for feedback, behavior, and survey research.

Feedback refers to the process of providing household-specific electricity use information back to the end users (individuals in a household) on some form of ongoing basis.[6] In most feedback programs, the goal is to prompt an individual or household to change their energy behaviors to save energy (either through efficiency, conservation, or demand response actions). EPRI's feedback type categorization is shown in Figure 1-1 below.

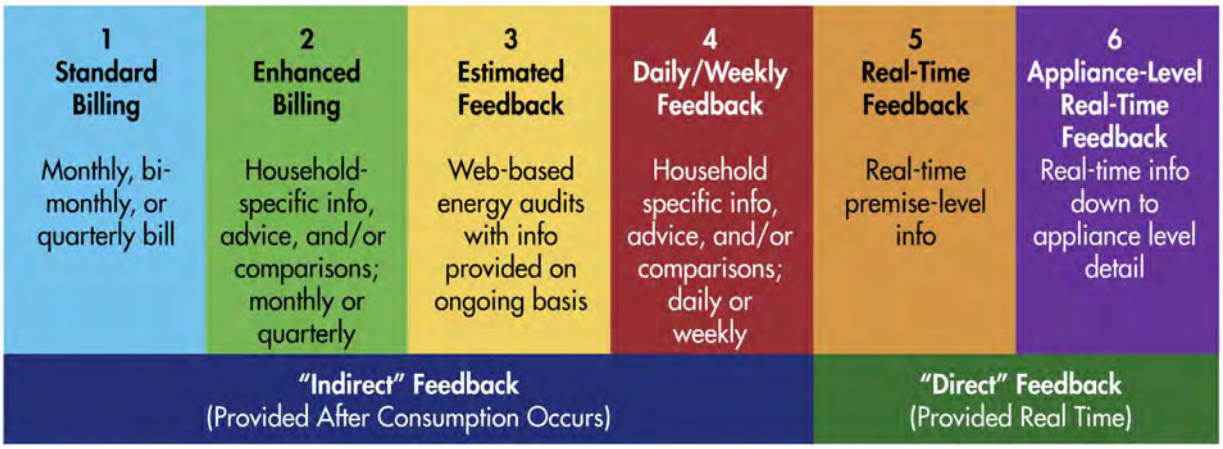

Figure 1-1
Feedback Delivery Mechanism Spectrum

Feedback categories 1-4 represent indirect feedback (where feedback is provided sometime after consumption occurs) and categories 5-6 represent direct feedback (where feedback is provided in real-time or near-real-time). Additionally, category 1 is generally universal—almost everyone receives a monthly or bi-monthly bill.

Behavior, as a term, has been subject to much debate in the energy industry. For the purposes of this report, we define "behavior" as any physical action taken on the part of feedback recipients that has an effect on the electricity consumption of the household.[7] In Table 1-1 we provide primary classes of behavior that are referenced throughout this report.

---

[6] While the focus of most feedback research is the residential sector, feedback can be provided to other sectors as well. For a review of commercial feedback interventions see EPRI. (2010). *Electricity-Use Feedback in the Commercial Sector: Examining the Potential for Building Occupant Behavior Change*. Palo Alto, CA: EPRI. Report No. 1021193.

[7] In addition to the aforementioned behaviors, measuring the precursors (awareness, knowledge, attitude, beliefs, perceived social norms, personal norms, sense of responsibility and values) to actual behavior change may help determine if behavioral programs are making progress toward the intended end result (behavior change). While we acknowledge that these factors are important to consider when understanding the full extent of a program's effects, this effort focuses on survey techniques to assess actual behavior changes. See Randazzo, K. V. & Peters, J. S. (2011). *Reconsidering What We Measure: Residential Decision-Making and Proposed Standard Questionnaire Items.* San Diego, CA: Presented to Southern California Edison and Pacific Gas & Electric.

Survey Research is the process of collecting information from a sample of a targeted population using structured, primarily close-ended questions. For the purposes of this report, we focus on traditional mail, internet and phone surveys. We do not discuss other methods of structured data collection, such as journaling or observation-based approaches, in this report; however, we acknowledge their potential value in quantifying effects of feedback programs.

Table 1-1
Primary Classes of Behavior

Behavior Type	Categories within Behavior Type	Description	Cost	Incidence/ Potential Frequency	Examples
Purchase Behavior	**Large Purchase Behavior**	Behavior that is defined by identifiable decision-making events, and often triggered by equipment breakdown or failure.[8]	High—installation requirement	5-20 years	Efficient HVAC Insulation Replace heat pumps Whole-house retrofit
	Moderate Purchase Behavior		Low-to-moderate with installation requirement	1-5 years	Weatherization such as window caulking Install motion sensor lighting
	Small Purchase Behavior		Low	6 months - 1 year (usually accomp. by larger quantity purchases)	Replace incandescent light bulbs with CFLs or LEDs
Usage/ Habitual Behavior	**Conscious**[9]	Discrete behaviors influenced by comfort, convenience, or maintenance of the household.	Little to no cost	Daily-Weekly-Monthly	Vaccum refrigerator coils Adjust the thermostat Clean/replace the air conditioning filter Clean dryer lint filter regularly Keep blinds or curtains drawn in the hot part of the day Close doors and windows when heat or cooling on
	Automatic	Behavior that is an automated cognitive process.[10] Consumers react almost subconsciously.	No cost	Daily or more frequently	Turn on/off lights, TV, etc. Using appliances when preparing meals

[8] Schwarz, N., Hippler, H., Deutsch, B., & and Stack, F. (1985). "Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments". *Public Opinion Quarterly*, 49, 388-395.
[9] Behaviors that are conscious for some might be automatic for others.
[10] Op Cit. (Schwarz, Hippler, Deutsch, & and Stack, 1985)
Op. Cit. (Randazzo & Peters, 2011)

2
OVER-ARCHING CONSIDERATIONS IN SURVEY STUDY DESIGN

Before beginning any survey research study, researchers have to carefully define the scope of the survey to ensure that the findings are relevant and defensible. This requires that researchers make a concerted effort to identify key pieces of information about the feedback intervention and effectively customize the survey research to the goals and unique qualities of the feedback intervention. To do this, researchers must consider three factors before developing a data collection method and survey instrument: (1) What is the intended use of the survey findings? (2)What type of information is being provided via the feedback? and (3) What types of behaviors are expected to change? In this section, we discuss each of these considerations and how the subsequent sections of this report address each of these questions.

What is the Intended Use of the Survey Findings?

Surveys can be used to support several different research objectives, and the survey study design and instrument should directly reflect these objectives. For example, the primary goal of a given survey may be to identify behavior changes for the purposes of an impact evaluation, or alternatively, the goal of the study may be to monitor actions for design and development—or both. While both goals require quantifying the behavioral effects of a feedback intervention, the levels of rigor required differ dramatically between surveys used for an impact evaluation compared to those used for program monitoring.

Impact evaluations require a defensible counterfactual, generated either through multiple measurements (such as pre-post studies or through a comparison group). However, this level of rigor may not be necessary or cost effective when collecting information on behavior change for the purposes of program monitoring or interim learnings. Section 3 helps researchers pick and choose among methods to ensure the study goals are reflected in the design phase.

What Type of Information is Being Provided?

In addition to developing the proper study design based on the objectives and types of feedback, researchers must determine which behaviors to target in a given survey instrument. For this reason, the type of information provided through feedback should guide the behaviors included in the survey. In Sections 4 and 5, we discuss how to prioritize and measure behavior changes based on the information provided and the implicit trade-offs necessary when choosing among them.

Unlike traditional rebate programs, feedback interventions motivate action by providing general and or directional information to participants. General information may be thought of as overall household feedback, such as average daily kWh, savings estimates, rates, or performance against a given benchmark (for example, the participant's or the participant's neighbor's usage). Directional information, conversely, provides participants with specific calls to action or "energy-saving tips" in an attempt to direct them to specific behavior change. A single intervention can offer both types of information. In the case of general information programs, the researchers must develop guiding principles for choosing the behaviors about which to collect information.

What Types of Behaviors Are Expected to Change?

While the information provided through feedback can help to determine the types of behaviors to measure, not all promoted behaviors may be feasible to study in the observation period,[11] nor should they all be measured the same way. In Section 3, we discuss timing of survey design and fielding for feedback efforts and the implications for analysis if the timing of data collection is not properly aligned with expected behavior changes.

For example, a feedback intervention may include tips which promote energy efficient central air conditioning purchases, but incidence in the population for such an action may be so low that the effect of such a promotion may not be measurable in the observation period without very large sample sizes. Further, actions like turning off the lights are so commonly cited that the researcher may not be able to detect increases in the exceptionally high incidence of the behavior. These considerations should factor prominently in deciding which actions to measure in a survey. In Sections 4 and 5, we discuss ways to and prioritize which behaviors to measure and the specific trade-offs that need to be made when measuring one behavior, versus multiple behaviors.

In addition, it is not likely that all behaviors will be affected in the same way through feedback interventions. In Table 1-1, we define five types of behaviors that, practically speaking, require different considerations and techniques when quantifying them through survey research. Purchase behavior requires different lines of inquiry than usage behavior. Purchase behaviors, for example, are more memorable and require a one-time decision, whereas usage behaviors are less remarkable and require multiple decisions over time, and may be best quantified in terms of the frequency and quality of the action. We discuss these considerations in detail in Sections 4 and 5 of this report**.**

[11] The observation period is defined as the timeframe in which the researcher is measuring the expected effects of the feedback program.

3
DETERMINING THE MOST APPROPRIATE SURVEY STUDY DESIGN

In Section 2, we describe three questions that researchers must consider when designing a survey research study to effectively scope and customize the study to the feedback intervention of interest. Past EPRI work has developed behavioral research design guidelines for feedback programs overall.[12] In this section, we revisit some of these concepts as they relate to survey research, data collection, and attribution analysis. Specifically, we discuss pre-post data collection approaches, the use of comparison groups, sampling, and the implications of different survey modes.

Why Study Design Matters for Survey Research

As with any type of data collection, proper research design is the foundation of survey-based studies. When faced with important research decision, if hasty design decisions are made based on perceived cost and expediency and de-prioritized accuracy, validity, and reliability, researchers can be left with data that are difficult to defend. To avoid this common pitfall, researchers should do the following before designing the survey:

- understand how the survey effort fits into the overall study design,
- establish the required level of rigor in attributing measured behavior change to the feedback intervention,
- understand the available methods for contacting individuals, and
- determine the sampling approach.

Understanding how the Survey Effort Fits into the Overall Study Design

In many cases, survey research is just one component of an overall research project for feedback programs. Typically, researchers will use survey research to complement an impact evaluation (which often include a billing or consumption analysis, or some other impact method),[13] or to inform program monitoring or process evaluations. In the best case, the feedback program of interest has an overall study design that may be leveraged for the survey research (such as a treatment and control group). However, in many programs, an experimental or quasi-experimental framework may not be in place prior to implementing the survey.

Researchers need to determine the required level of rigor for their survey efforts, irrespective of any pre-existing study frameworks. For instance, if the goal of the survey is to gain general customer opinions on a feedback intervention, a participant and control group study may be unnecessary even if an experimental infrastructure is in place. Conversely, if the goal is to determine

[12] Op. Cit. (EPRI, *Guidelines for Feedback Pilots,* 2010.)

changes attributable to feedback in support of an impact evaluation, a comparison group should be carefully considered even if such a group was not created in the program design. To identify the most appropriate survey study design, researchers should determine the required level of rigor for their efforts.

Choosing the Right Study Design to Support Desired Levels of Rigor

When considering the intended use of the data collection methods, it is necessary to consider the required level of objectivity and rigor for the study. One of the key considerations is the level and extent of information needed to establish program attribution. Attribution is established by determining a counterfactual, e.g., what would have occurred in the absence of the program. Any observed differences between the counterfactual and participant group is, generally speaking, considered attributable to the feedback intervention.

Every data collection method has a varying degree of objectivity when establishing attribution, and, depending on the intended application of the data and budgetary constraints, researchers must carefully choose among them. Wherever possible, researchers should begin with the goal of establishing a counterfactual or reference point to establish whether the observed behavior changes are due to the feedback program rather than natural occurrences. There are four methods for gathering a counterfactual through survey data: (1) using a control or comparison group (an equivalent population that did not receive feedback treatment); (2) using multiple measurements of the participant population (i.e., pre- and post- treatment); or (3) combining approaches (1) and (2) together; and (4) using self-report to establish changes since the feedback intervention.

Overall, the most rigorous approach involves using both a control or comparison group as well as pre-post measurements (i.e., method 3). However, in the real world, researchers can often be left to measure program effects as much as one to two years after the program launched with limited input into program design. In these situations, researchers lose the ability to establish a baseline or a treatment or comparison group. As a result, researchers often rely on participants to report whether they made a change in response to the feedback treatment (post-only self-report). This is a much less desirable, but more common, method for establishing causality.

Below we briefly describe various survey design methods and demonstrate the differences in the types of data that each one provides using one “behavioral” example (Exhibit 3-1). In addition, we briefly discuss the survey mode and sample.

Use of a Comparison Group for Survey Efforts

In survey research, the use of a comparison group allows researchers to observe change between two equivalent populations where one population is treated with a feedback intervention and one is not (See Figure 3-1: A and Figure 3-1: C). Differences measured between these groups at the time of the interview are assumed to be due to the intervention of the feedback mechanism.

When using the comparison group, survey questions should be identical for both populations to allow for this comparison. By looking for statistical differences between their survey responses, researchers can determine the effects of the feedback intervention and thus do not need to ask the treated group to self-report the influence of the feedback intervention on their actions.

It is important to note that developing a comparison group is, in itself, a research task that requires sophisticated statistical techniques. As outlined in past EPRI work,[14] the most rigorous method involves control (and treatment) groups developed for randomized control trials (i.e., the pool of potential participants are randomly assigned to treatment and control groups at the same time). That said, other methods to develop comparison groups can be used if this is not possible. In these situations, it is important to note here that concerns center on determining the equivalency of the treatment and comparison group. Namely, without random assignment, it is difficult to ensure that the treated group is truly comparable to the comparison group in all factors that might affect the variety of behaviors covered through feedback interventions. For this reason, it is important that comparison groups are identified by a seasoned statistician familiar with the utility industry to ensure that comparison groups are as equivalent as possible given the limitations of the study and available data on the treatment and comparison populations. Again, we refer readers to the past work for more information.[15]

"When dealing with feedback-based programs, without random assignment, it is difficult to ensure that the treated group is truly comparable to the comparison group in all factors that might affect the variety of behaviors covered."

Pre- and Post- Treatment Survey Efforts

When done correctly, a pre- and post-treatment survey effort allows researchers to measure change over time in a given population (see Figure 3-1: A and Figure 3-1: C). This method first establishes baseline behaviors prior to the treatment, and then measures the intervention's effects by conducting a follow-up survey after the intervention (post study). The differences between the baseline (pre) and the follow-up survey (post) are generally considered attributable to the feedback intervention.

When planning a pre- and post-treatment survey effort to measure behavior change, the most important concern is "timing," i.e., both the timing of when you are fielding (e.g., your observation period), and the timeframe stated in the questions. In this type of study, the questions must be designed to account for both the overall timing of the intervention and the fielding times of the survey effort. For example, both the pre- and post-treatment questions should ask about behaviors at a single point in time, or over a similar timeframe. Notable mistakes with this type of research include asking about a dissimilar timeframe, or comparing two seasons that may be dissimilar if there is a seasonal pattern to usage. Timing, therefore, needs to be carefully considered. Exhibit 3-1 further illustrates this point.

[14] Op. Cit. (EPRI, *Guidelines for Feedback Pilots,* 2010.)
[15] Ibid.

GOOD:

Asking about a behavior at a single time period, pre-post

(Pre-treatment, June Year 1) Q1. Thinking about the most recent time you were home, at what temperature (in degrees Fahrenheit) did you set your thermostat?

(Post-treatment, June Year 2) Q1. Thinking about the most recent time you were home, at what temperature (in degrees Fahrenheit) did you set your thermostat?

Asking about behavior across a time period, pre-post

(Pre-treatment, asked in June Year 1) Q2. In the past year, have you had a tune-up for your central air conditioner?

(Post-treatment, asked in June Year 2) Q2. In the past year, have you had a tune-up for your central air conditioner?

BAD:

Asking about a behavior at a single time period, pre-post, but measuring during different season

(Pre-treatment, June Year 1) Q1. What temperature (in degrees Fahrenheit) did you set your thermostat to the most recent time you were not at home?

(Post-treatment, January Year 2) Q1. What temperature (in degrees Fahrenheit) did you set your thermostat to the most recent time you were not at home?

Asking about behavior across a time period, pre-post, but not a comparable timeframe:

(Pre-treatment, June Year 1) Q2. In the past year, have you had a tune-up for your central air conditioner?

(Post-treatment, June Year 2) Q2. In the past three months, have you had a tune-up for your central air conditioner?

Exhibit 3-1
Example Survey Questions Illustrating Differences in Timing

While pre-and post-treatment surveys offer a better understanding of pre-treatment conditions and changes over time, pre- and post- treatment surveys on their own (without a comparison group) do not account for naturally occurring effects, such as economic effects or national reductions in CFL sales due to federal legislation or economic effects (see Figure 3-1: B).

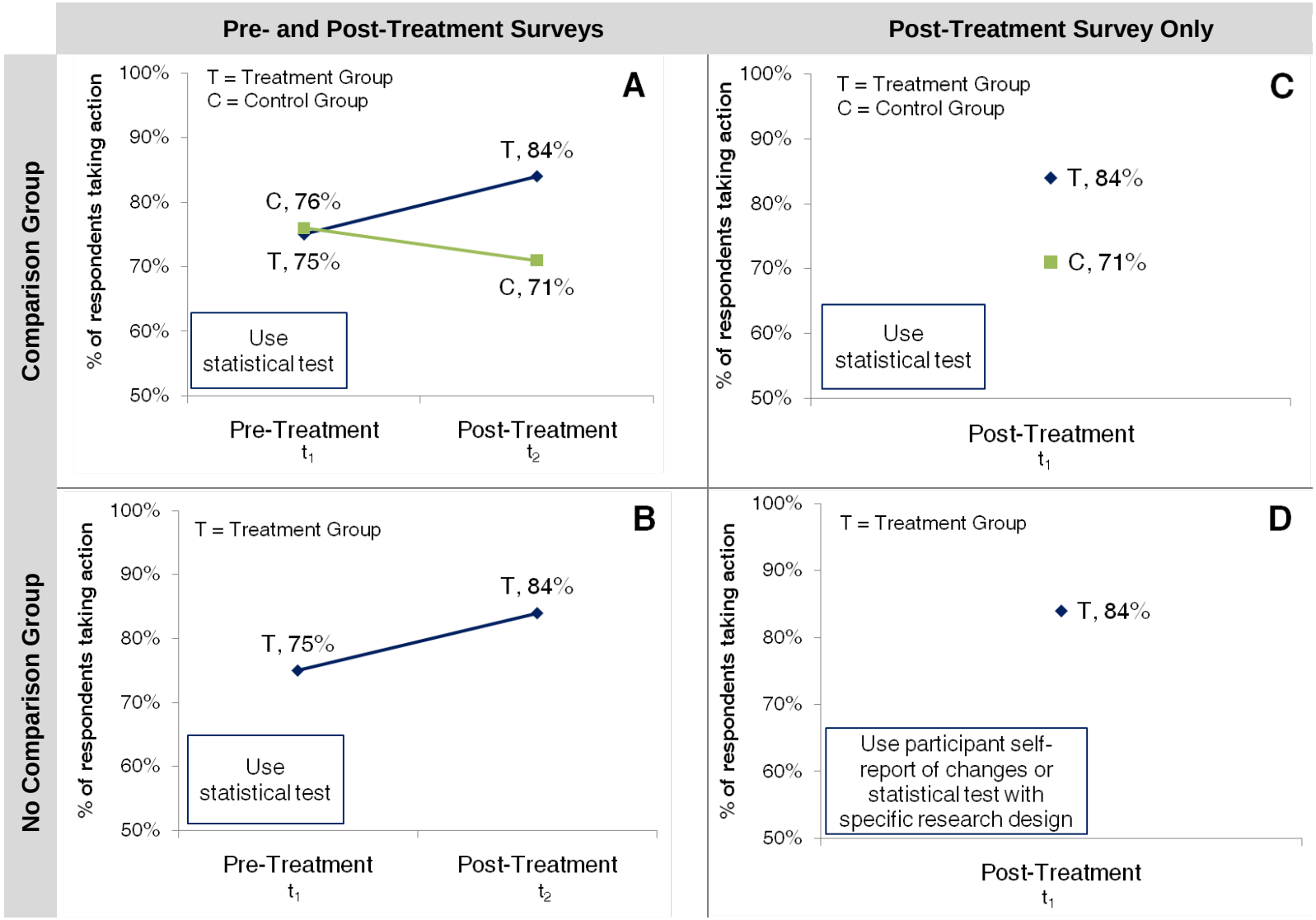

Figure 3-1
Examples of Survey Study Designs

*Note the differences between T_2 and C_2 must be statistically significant at a pre-determined criterion level.

**Note that since this uses a comparison group only a minimal difference is expected for the pre-treatment period.

Post-Treatment Only Survey Efforts

Generally, post-treatment only survey efforts can be used to make generalizations about the population at a certain point in time; however, when trying to use post-treatment only surveys to understand changes as a result of a feedback mechanism, the researcher must rely on self-reported data to understand the effects (see Figure 3-1: D). This is the least reliable method, but at times, may be the only option given budget limitations and/or the timing of when the research started. Often times, survey researchers are brought in after the fact, and thus are not able to establish a baseline or comparison group to understand the counterfactual.

When asking about behavioral change, post-only surveys need to ask about two points in time, or directly ask about changes in behaviors.

EXAMPLE:

Two points in time:

Q1. BEFORE you started getting information from your home area network, did you set your thermostat to recommended set points (e.g., 78 for cooling/68 for heating)?
Q2. AFTER you started getting information from your home area network, did you set your thermostat to recommended set points (e.g., 78 for cooling/68 for heating)?

Directly asking about changes:

Q1. Since you started receiving information from your home area network, have you started to set your thermostat to recommended set points (e.g., 78 for cooling/68 for heating)? (NOTE: Include option for "I was already taking this action.")
Q2. Since you started receiving information from your home area network, have you had a tune-up for your central air conditioner?

Exhibit 3-2
Examples of Post-Treatment Only Survey Questions for HVAC Behaviors

The survey questions used in a pre-post study differ from those used in a post-only study. In the example below, it can be seen that the pre-post study survey questions asks about an action using a single point in time. In a post-only survey question, two time references are used—before and after the treatment. To further illustrate these differences, we provide examples below in Exhibit 3-3 and Exhibit 3-4.

PRE-POST SURVEY QUESTION:

Asking about a behavior at a single time period, pre-post

(Pre-treatment, June Year 1) Q1. Thinking about the most recent time you were home, did you turn off your TV when it was not in use?

(Post-treatment, June Year 2) Q1. Thinking about the most recent time you were home, did you turn off your TV when it was not in use?

Exhibit 3-3
Examples of Pre-Post Survey Questions

POST-ONLY SURVEY QUESTION:

Two points in time:

Q1. BEFORE you started getting information from your home area network, did you turn off your TV when it was not in use?
Q2. AFTER you started getting information from your home area network, did you turn off your TV when it was not in use?

Directly asking about changes:

Q1. Since you started receiving information from your home area network, have you started to turn off your TV when it was not in use? (NOTE: Include option for "I was already taking this action.")

Exhibit 3-4
Examples of Post-Only Survey Questions for Electronics Behaviors

The primary distinction of post-only methods is that they require the respondent to self-report attribution to the program. While there are generally accepted practices in the industry for asking about attribution to programs for standard incentive-based programs, these are not appropriate to measure behavior changes from feedback mechanisms because (1) standard attribution batteries are generally asked for *each* behavior and feedback-based programs may promote numerous behaviors; and (2) the information provided by the feedback mechanism may have educated consumers causing a cognitive change—there isn't just one action (like receiving a rebate) that can be easily recalled. Sample questions to measure self-reported attribution from feedback mechanisms are described in Section 4.

Data Collection Options based on Available Methods

In addition to determining the "design" of the survey effort (post-only, use of a comparison group, etc.), the research should also consider the best fielding option or "mode". Generally, the mode is determined by available customer contact information, but feedback programs should seek to gather as much contact information as possible (without undue burden) to enable choice for any future data collection effort. There are four primary modes of data collection relevant to this discussion:

- Self-administered mail surveys
- Self-administered internet/email surveys
- Interviewer-administered phone surveys
- Multi-mode surveys (combination of self- and interviewer-administered surveys)

Multi-mode (e.g., online and telephone) is generally the best option to reduce self-selection and other biases, as this approach incorporates the benefits of both modes; however, multi-mode requires an additional level of review to ensure that responses from both methods can be combined in the analysis phase. For example, in an internet survey, respondents almost always "see" possible response categories. To ensure equivalency with telephone survey results, all pre-coded responses need to be read out loud by the telephone interviewers to reduce mode-based bias. However, for phone surveys, this often results in an even longer and more burdensome survey.

While multi-mode is ideal, online efforts are generally most appropriate if the feedback intervention is such that web or email communication has already been established with participants. Given that it may be of interest to investigate several behaviors via the surveys, online efforts also allow for more extensive questions (that is, covering a broader range of behaviors), and they can be fielded inexpensively.

However, if the study also includes a comparison group, online data collection may not be an option if email addresses are not available for the comparison group. Moreover, additional cognitive testing may be required to ensure that the questions are measuring responses effectively since online survey efforts cannot be monitored for cognition or respondent's comprehension of the question, an implicit value in interviewer-administered surveys.

Each of the modes used to field surveys have certain advantages and disadvantages. These advantages and disadvantages should be considered when establishing the survey study design. Table 3-1 summarizes the advantages and drawbacks for researchers' consideration.

Table 3-1
Data Collection Options[16]

Variable	Telephone Interview	Mail Survey	Web Survey	Multi-Mode (Telephone/ web)
Cost of fielding	$$	$$	$	$$$
Cost of data processing	$	$$$$	$	$$
Speed to field	Fast	Slow	Moderate	Moderate
Response rate[17]	Moderate	Low	Moderate-Low	Moderate-High
Ability to allow for…				
Lengthy questionnaires	Moderate	Moderate (expensive to print)	High (allows for more questions to be covered quickly)	Moderate
Sensitive questions	Moderate	High	High	Moderate-High (not comparable)
Lengthy answer choices	Low	High	High	Low

[16] Owens, L. K. (2005). *Introduction to Survey Research Design*. Chicago: University of Illinois at Chicago Survey Research Laboratory.
Roberts, C. (2007). *Mixing modes of data collection in surveys: A methodological review*. London: Briefing Paper for The Economic and Social Research Council, National Centre for Research Methods.
Dillman, D. (2007). *Main and Internet Surveys – The Tailored Design Method* (2nd ed.). New Jersey: John Wiley and Sons Inc.
[17] Response rate will of course vary depending on many factors, such as instrument design, sample population, etc.

Table 3-1 (continued)
Data Collection Options

Variable	Telephone Interview	Mail Survey	Web Survey	Multi-Mode (Telephone/ web)
Complex skip patterns	Moderate	Low (cannot ask people to skip as easily as when programmed)	High	Moderate
Complex questionnaire concepts	Moderate	High	High	Moderate (default to lowest)
Social desirability bias	Moderate	Minimized	Minimized	None-Moderate
Quality of responses	Moderate-High	Low (cannot ensure responses)	Moderate-High	Moderate-High
Representation concerns[18]	May exclude households using cell phones only (or double-count households with both cell phones and landlines)	Low response rates	Excludes households without computers or Internet access	Most representative, but may have variation in responses between modes

Sample Selection

Because researchers do not generally have sufficient budget or time to administer surveys to an entire population, sampling is used. A good sample will represent the population well and serve as a tool to find out about more about that population. A detailed discussion of sampling is beyond the scope of this research, but we again refer the reader to EPRI's past work.[19] Here, we focus on the primary considerations for survey research.

The sample size indicates the number of people that the survey successfully reached to provide data and is typically the number that most researchers look at to determine whether they feel confident in the survey results.

There are a number of considerations in sample sizes discussed in the aforementioned EPRI report; however, here we point out one particular consideration for programs looking at changes in behaviors related to feedback interventions. Generally, sample size needs to differ based on whether the researcher need to estimate proportions, ratios, or means with a certain degree of accuracy. It also depends on the size of the effect that the researcher expects to see. Notably, effect sizes can vary significantly across various behaviors; and for some behaviors with high prevalence in the population, it will be difficult to determine change without extremely large

[18] This is largely dependent on the population and sample frame, but these are the general issues/concerns with each mode.
[19] Op. Cit. (EPRI, *Guidelines for Feedback Pilots,* 2010.)

sample sizes. However, simply having a large sample of completed surveys does not guarantee valid results. We recommend that an experienced sample designer be used for all surveys to be sure that the appropriate people are included in the sample frame, and that the sample size will provide the highest level of precision and confidence for the resources allocated to the project.

Selecting the Best Survey Design

The specific types of behaviors changed as a result of feedback mechanisms have important policy and planning implications. For this reason, researchers and program implementers should seek out the most rigorous and conservative methods for establishing program effects from feedback interventions. If causal relationships are not sufficiently characterized to support confident and reliable predictions of a feedback program's effects (as compared to standard programs), research findings may (wrongly) serve as the basis for future program planning. Thus, over or under-representing the savings potential of these programs can have significant consequences for future programs.

Survey study design should be considered with this in mind, and researchers should therefore default to a method that provides the greatest objectivity within the context of the overarching research design, which usually also includes a consumption data analysis. Wherever possible, researchers should seek to establish comparison groups or use multiple measurements (ideally using a pre-post treatment-control group design) to determine whether the feedback program had an effect on the participants' behavior changes (causality).

4
THE COMPLEXITIES IN MEASURING BEHAVIOR CHANGE

While asking about behavior change may seem like a simple undertaking, it is possible that electricity use feedback may affect a wide range of behaviors—from conservation actions to high-cost, whole-house retrofits. Given cost limitations and the limited patience of the respondents, no single study can measure all potential behavior changes resulting from feedback programs. Thus, the range and diversity of potential behavioral effects makes it difficult to identify "best" behaviors to measure and the "best" study design techniques to collect this data.

Feedback programs have the potential to impact a wide range of behaviors, from purchase behavior (high, medium, and low cost) to usage behaviors (conscious and automatic). Table 4-1 and Table 4-2 detail the sheer range of actions that could be affected by a feedback program.

While these lists are long, they are not exhaustive and researchers must make trade-offs between enhanced precision/bias reduction and collecting data on all behaviors of interest. In this section, we provide guidelines for prioritizing which actions to measure and how to carefully quantify respondents' changes.

Table 4-1
Examples of Potential Purchase Behaviors to Measure through Survey Research

High Cost Actions	
1.	Central air conditioning unit
2.	Room or wall air conditioning unit
3.	Clothes washing machine
4.	Clothes dryer
5.	Dishwasher
Medium Cost Actions	
6.	Television
7.	Printer
8.	Computer
9.	Video game console
10.	Outdoor light fixtures
11.	Indoor light fixtures
12.	Boiler
13.	Furnace
14.	Refrigerator
15.	Pool
16.	Pool Pump
17.	Attic, ceiling or wall insulation
18.	Programmable thermostat
19.	On-demand or tankless water heater

Table 4-1 (continued)
Examples of Potential Purchase Behaviors to Measure through Survey Research

Low Cost Actions	
20.	Weather stripping or caulking around windows and/or doors
21.	Energy Star light bulbs or compact fluorescent lights, also known as CFLs
22.	Low-flow shower heads
23.	Faucet aerators
24.	Water heater wrap
25.	Insulated outlets and/or light switches
26.	Fluorescent lights that are not compact (i.e., fluorescent bulbs that are longer and thinner than CFLs)
27.	Motion sensors (for lighting)
28.	Lighting timer(s)
29.	Task lighting
30.	Window shades, window insulation or window quilts

Table 4-2
Examples of Potential Usage Behaviors to Measure through Survey Research

Usage Behaviors	
1.	Laundry water temperature
2.	Method of drying laundry (e.g., hang laundry to dry, machine dry with moisture sensor, machine dry with timer)
3.	Washing machine loading (e.g., fully loaded, partially loaded)
4.	Dishwasher loading (e.g., fully loaded, partially loaded)
5.	Turn on/off lights in unoccupied rooms
6.	Use task lighting
7.	Use lighting timer
8.	Turn off outside lights by day
9.	Turn off computers at night/when not in use
10.	Put computer(s) to sleep
11.	Turn off TV(s) when not in use
12.	Turn off video game console(s) when not in use
13.	Switch off power strips or unplug devices when not in use (chargers, TVs, stereos, etc)
14.	Lower window shades, insulation or quilts
15.	Take short showers
16.	Use a portable window fan
17.	Maintain your heating and cooling system
18.	Change the furnace filter
19.	Clean the boiler water
20.	Reduce the water heater temperature
21.	Clean or replace air filters
22.	Clear the area around vents
23.	Make sure refrigerator seals are tight
24.	Clean refrigerator coils
25.	Check refrigerator temperature
26.	Unplug a second refrigerator for weeks to months at a time
27.	Set the thermostat to recommended set points (e.g., 78 for cooling/68 for heating)

Which Behaviors Should the Survey Measure?

In an ideal world, researchers would have unlimited budget and respondents would have unlimited patience; in reality, researchers have to choose the behaviors to measure with a survey if they are to measure them well. Similar to selecting the appropriate survey data collection method, researchers must carefully consider the feedback program and the intended use of the study findings in their analysis. This can be done by revisiting the key questions we outlined in Section 2.

- **What is the intended use of the survey findings?** Before implementing a survey effort, researchers should determine how the survey findings will be used. If the goal of the study is to determine energy saving potential, then researchers should default to collecting detailed information on those behaviors that are likely to have occurred in the observation period and/or that have the potential for the greatest energy savings. Conversely, if the goal is to determine the potential range of actions that may occur due to feedback, researchers may want to systematically represent behaviors that are different (despite how much energy they save) in order to illustrate the breadth of the programs' effects.
- **What type of information is being provided?** In addition to considering the goals of the study, another approach to narrowing the scope of behaviors is to determine which actions the program promoted during the observation period. By targeting these actions, researchers are focusing on the actions that the program was aiming to affect, thus aligning with the program theory. However, in many feedback programs, general information is provided, such as usage and rates, and it is up to the participant to identify actions that will reduce their usage and/or rates. In these cases, the researcher should carefully consider which actions could have been affected during the observation period based on program design.
- **What types of behaviors are expected to change?** In addition to considering the goals of the program and the actions promoted, researchers should also consider the types of actions that are likely to have occurred during the observation period. For instance, if the survey is collecting data three months after participants are first exposed to feedback, the researchers should focus on usage behavior and low-cost actions. Due to low incidence of purchase in the population, high-cost actions are unlikely to be affected, or at the very least detected, in a survey at this point in time. However, surveys administered after a year of exposure are more likely to capture high-cost purchases.

In addition, researchers should consider the season(s) and corresponding purchase patterns when deciding which actions to measure as these factors influence the likelihood that these actions occurred. For instance, it does not make sense to inquire about thermostat settings in winter months during the summer season, or to discuss planting shade trees in the middle of the winter.

Researchers must decide the behaviors that are a priority for the purpose of the study and make trade-offs between enhanced precision/bias reduction and collecting data on all behaviors of interest. For this reason, researchers have to prioritize which behaviors to measure weighing the pros and cons throughout the decision-making process. Once this is done, the survey instrument must then be crafted to effectively quantify changes in these behaviors. This is discussed in detail below.

Identifying the Appropriate Level of Detail for Each Behavior of Interest

Due to cost and time restrictions, researchers have to choose between collecting detailed information on a select set of behaviors or gaining high-level information on multiple behaviors. Each survey question represents valuable "real estate" and researchers have to choose among behaviors of interest if they want to get specific information, such as thermostat set points or number of efficient light bulbs. If a particular behavior is of a high priority, then researchers must take the time to ask the appropriate questions to avoid getting false levels of precision due to poor recall. Below are some examples of how very specific information about a certain action may be obtained if researchers are interested in greater levels of detail for specific actions.

EXAMPLES:

Thermostat Setting: Q1. We would like to know more about your thermostat setting. Could you please go and take a look at the setting and tell us the exact temperature you have it set to?

Q2. Does this set point change for times when you are typically out of the home? If yes, what does it change to?

CFLs: Q1. We would like to know more about CFL usage in your household. While we are on the phone, I would like you to walk through each room of your home and count the number of screw-in light bulbs. Please take scratch paper and a pencil with you to help keep count. For each room, please tell me the total number of fixtures and the number of CFLs you have installed.

HVAC SEER Rating: Q1. We would like to know more about your central air conditioning system. To help us determine the efficiency level of your system, we need you to provide us with the model number. Please take a moment to locate your owner's manual or the plate on your system with this information and provide the model number in the space below.

Exhibit 4-1
Examples of Levels of Detail Required for Different Household Behaviors

However, as explained before, researchers cannot ask detailed questions about all behaviors. In the question examples that follow, a less detailed but more efficient approach for capturing behavior changes is used where the researcher is able to gather more information about multiple behaviors rather than about just about a selected few. For instance, the above example asks the participant to count the exact number of CFLs present in the house, whereas a more efficient example would be asking the participants whether they have installed CFLs in their home and then follow-up with a battery of questions (see Appendix B).

EXAMPLES:

Specific Information: Q1. We would like to know more about CFL usage in your household. While we are on the phone, I would like you to walk through each room of your home and count the number of screw-in light bulbs. Please take scratch paper and a pencil with you to help keep count. For each room, please tell me the total number of fixtures and the number of CFLs you have installed.

More Efficient Approach: Q1. I am going to list items or equipment that might be in your home. Does your home have CFLs? *(This is then followed by a battery of questions—see Appendix B)*

Exhibit 4-2
Example of More Efficient Wording of Behavior Questioning

Developing an Effective Instrument to Measure Changes in Behavior

For each behavior of interest, researchers must design an instrument that at once *quantifies potential behavior changes* across all categories while also collecting information needed to *attribute* these changes to the feedback intervention. In this section, we provide guidelines for researchers to effectively design an instrument to do both.[20] We then provide example batteries of survey questions to further illustrate the level of detail and attention needed to effectively collect this information.

In this section, we provide example survey questions to measure behavior changes using a pre-post or comparison group format. As you will note throughout, the question wording does not ask directly whether the feedback intervention caused the action, because the comparison pre-post t_2-t_2 or participant to comparison group T-C measures the influence of the program (see Figure 3-1).

Quantifying Potential Behavior Change

Once researchers have established which types of behaviors to measure, they must then determine, specifically, which actions respondents have taken. This requires determining:

1. The presence or occurrence of the behavior (e.g., the purchase of an air conditioning unit or turning off the lights).
2. The level of energy savings of the behavior (e.g., an ENERGY STAR air conditioning unit, or how regularly the participant turns off the lights).

At first pass, it may seem as simple as asking a direct question such as, “Did you purchase an energy efficient air conditioning unit?” or “Do you turn off the lights when you leave a room?” However, there are a number of implicit assumptions in a single question that can obscure or bias the information collected in a survey instrument. When we unpack the former example question,

[20] It is important to note here that if researchers are interested in quantifying the energy saving impacts of attributable behavior changes, characteristics of systems and more detailed building information would be required.

"Did you purchase an energy efficient air conditioning unit?" a number of implicit assumptions become clear. This phrasing assumes that:

- The respondent has the ability to take the action,
- The respondent is accurately identifying the equipment as "energy efficient," and importantly,
- The reported action has a positive effect on energy consumption (i.e., reduces energy usage) if respondent answers "yes" rather than, say, adding a new unit to the household.

Taken at face value, a simple "yes" response would appear to indicate a reduction on the respondent's energy usage, but this may not be the case. However, for each type of behavior (purchase versus usage), there are a number of data points that should be collected to better establish the presence and level of energy savings of each behavior type (purchase and usage). Below we provide guidelines for each.

Purchase Behaviors

Presence of the action: When establishing the presence of the action, the researcher must determine whether the respondent is likely to take the action. In most cases, this is largely a function of their control over their household and the existence of the equipment of focus.

1. *Respondent has the ability to take the action:* In the case of purchase behavior, it is important to establish whether the respondent can take a given action. In many cases, renters are neither able nor likely to make large changes to their building or appliances. For this reason, it is not prudent to ask them question that typically apply to owners only. From an analysis standpoint, it may also not be appropriate to compare these customers to owners as they are not "eligible" to take action. Owners in multi-family dwellings may have similar restrictions.
2. *Presence of existing equipment:* Once you establish the customer's control over their home equipment, it is important to establish whether or not the customer has any of the equipment of focus. Once this is established, researchers can then follow-up to determine if the equipment is "new" and "efficient." To determine whether it is efficient, visual aids, such as the ENERGY STAR logo, can be used to eliminate confusions.

   In addition, these types of questions allow the survey to be programmed so that all subsequent questions reflect this knowledge. For instance, once it is established whether or not a participant has a programmable thermostat, subsequent questions can focus on either manual adjustment behaviors or programmed set points.
3. *New action taken since program launch:* After determining that the equipment is in the home, researchers can then establish whether or not this equipment was purchased since the program launch. By doing so, researchers then identify which actions the program had the potential to affect. In all cases, the observation period (the time in which you are expecting to observe effects) should be carefully specified. See examples provided in Section 3 for guidelines for instrument development for pre-post and comparison group approaches.

Level of energy savings of the action: After identifying which actions may have been affected by the program, the researchers need to determine whether the purchase behavior has increased or decreased the respondent's energy usage.

4. *Efficiency level:* Some purchase behaviors, such as weather stripping or CFL purchases, are inherently efficient. However in these instances, researchers may want to consider the amount of the product in a home (number of CFLs, thoroughness of weather stripping). Equipment, on the other hand, requires that researchers establish the level of efficiency of the action. In most cases, researchers should avoid terms like "energy efficient" that may be interpreted as relative, and default to using concrete benchmarks, like ENERGY STAR labels or SEER ratings. In the case of specific energy labels visual aids or carefully worded explanations are required.
5. *Additional or replacement*: Finally, if researchers are interested in determining the impact of these purchases on energy usage, researchers need to inquire as to whether or not the purchase was a replacement or an additional unit.

In the following example, we provide a battery of questions that attempt to gather this type of information from respondents. In this case, we use a single example, an energy efficient air conditioning unit, although these questions may be relevant to other purchase behaviors as well. The questions below are designed to accommodate multiple actions in a single instrument.

Question 1: (Respondent has the ability to take action) Do you or members of your household own this home or do you rent?

1. Own/Buying

2. Rent/Lease

[SKIP Question 2-5 if Question 1 = Rent/Lease]

Question 2: (Confirm presence of the behavior) Does your home have any of the following items:

	Yes (1)	No (2)	Don't Know (98)
Central air conditioning unit			

[SKIP IF Question 2=No]

Question 3: (Confirm the behavior occurred in the observation behavior) Please indicate whether your Central air conditioning unit was purchased or installed in the past year.

	Yes (1)	No (2)	Don't Know (98)
Central air conditioning unit			

[SKIP IF Question 2 = No]

Question 4: (Determine the level of energy savings) Please indicate whether the Central air conditioning unit you installed in the past year was an additional unit or replacement unit.

	Additional Unit or System (1)	Replaced an older model (2)	Don't Know (98)
Central air conditioning unit			

[SKIP IF Question 2 = No]

Question 5: (Determine the level of energy savings) Was the Central air conditioning unit you purchased in the past year an ENERGY STAR Product?

	Yes (1)	No (2)	Don't Know (98)
Central air conditioning unit			

Exhibit 4-1
Examples of High-Cost Purchase Behavior Battery

Usage Behaviors

Unlike most purchase behaviors, usage behaviors have degrees of engagement based on the frequency or intensity of an action. Below we describe how the primary considerations for measuring purchase behaviors differ when considering usage behavior.

Presence of the action: In the case of usage behaviors, researchers should consider whether the action is equipment-dependent or not. If it is equipment dependent, then the researcher needs to establish whether the respondent has the equipment prior to inquiring about usage behavior (e.g. seasonal set points on programmable thermostats). However, unlike purchase behavior, usage behaviors typically occur much more frequently, from seasonal actions to daily behaviors. In this respect, the researcher needs to establish the presence of the behavior during the survey and then determine if the behavior had the potential to be affected by the feedback program (namely, did it start, increase, or decrease in frequency during the feedback period).

1. *Respondent has the ability to take the action:* In the case of usage behaviors, researchers must establish whether the behavior is dependent on situational constraints, such as the presence of certain equipment or homeownership, before inquiring about action.
2. *Presence of the behavior:* Once researchers have established that the respondent's behavior has the potential to be affected (based on #1), the researcher then must establish whether the behavior of interest exists. In the case of usage behavior, researchers may consider asking, for example, "Do you turn off your computer every night before bed?"
3. *Action has changed since program launch:* For usage behaviors, researchers need to establish whether the action "began" or was initiated for the first time during the treatment period. However, in the case of actions for which their frequency can change, researchers have to establish whether or not the action, if not initiated during the observation period, changed during the observation period. For this reason, if the respondent replied "Yes," to a question posed in #2 above, researchers should still follow up with a question that establishes whether the action was newly initiated or changed in frequency during the observation period (see point #4 below). See examples provided in Section 3 for guidelines for instrument development for pre-post and comparison group approaches.

Level of energy savings of the action: Similar to purchase behaviors, usage behaviors have varying degrees of efficiency or savings potential that can affect the level of efficiency or the frequency of behavior.

4. *Efficiency level or changes in the frequency of the behavior:* These can be thought of in two primary ways: (1) setting-related actions, such as temperature settings or (2) the frequency of the actions, both relative changes (increases or decreases) or time-specific (seasonal, monthly, daily).
    a. *Settings:* Many appliances have settings that can affect how much energy a given appliance uses. These can range from thermostat set points, the temperature settings of a water heater, sleep settings on a computer, or display settings on a television. While these actions directly affect equipment, they are usage behaviors and are contingent on the conscious actions of end users.

b. *Changes in frequency of the action:* In addition to setting behaviors, customers may change the frequency of their behaviors. For example, program participants can respond to feedback by more-consciously increasing the frequency of an existing behavior, maintaining the level of behavior, or decreasing the behavior.

In the following example, we provide a battery of example questions for the behavior "turning off your computer at night."

Question 1: (Respondent has the ability to take action) Do you own a computer?

1. Yes
2. No

[SKIP Question 2-4 IF Question 1 = No]

Question 2: (Confirm the presence of the behavior) Please indicate whether you regularly take any of the following actions in your home?

	Yes (1)	No (2)	Don't Know (98)
Turn off computer at night			

[SKIP IF Question 3 = No]

Question 3: (Confirm the behavior occurred in the observation period) Did you starting taking this action in the past year?

	Yes (1)	No (2)	Don't Know (98)
Turn off computer at night			

[SKIP IF Question 2 = No and Question 3 = Yes]

Question 4: (Confirm the behavior occurred in the observation period, determine the level of energy savings) Please indicate whether you increased or decreased the frequency with which you took this action in the past year compared with previous years.

	Increased Frequency (1)	Decreased Frequency (2)	No change in frequency (3)	Don't Know (98)
Turn off computer at night				

Exhibit 4-3
Examples of Usage Behavior Battery

Determining if Observed Changes are Attributable to the Feedback Program

In addition to measuring changes that occurred during the treatment period, researchers must also establish whether or not the actions taken are *attributable* to the feedback program. To do this, researchers must answer two main questions: (1) Are any of the observed changes attributable, at least in part, to participation in other programs? (2) Did the feedback mechanism *cause* these observed changes?

For the first question, the survey instrument should attempt to establish the influence of other programs, such as rebates, tax incentives, give-aways, or audits. However, establishing causality is determined either through the design of the study (e.g., through the use of a pre-post approach or comparison groups, discussed in Section 3) or through establishing a self-reported level of influence battery. We discuss how to measure the influence of other programs and guidelines for collecting self-reported influence.

Measuring Cross-Program Participation

Because feedback programs promote a wide range of actions, including, in many cases, other federal or utility run programs, it is important to determine, to what extent, the feedback program has channeled customers into other programs. This is important for two reasons: (1) to demonstrate the influence of feedback programs on the overall program portfolio; and (2) to identify risks of "double-counting" savings. Accounting for double-counting is more commonplace for impact evaluations (cross participation of feedback and rebate program participants) and policy makers and evaluators may expect to see such questions in the survey instrument. However, the actual task of identifying double-counted savings typically occurs through cross-referencing participant databases.

To establish other program influence, the survey instrument should attempt to measure respondents' use of rebates, tax incentives, give-aways, or audits. The most rigorous way to do this is through the use of a comparison group, as pre-post and post-only data collection methods cannot capture such effects.

Below, we provide an example of how the former air conditioning example could be expanded if the researchers want to determine the potential influence of other programs.

Question 6: (Determine other program influences) Did you receive a rebate or a tax incentive for the central air conditioning unit you installed or purchased in the past year?

	Rebate (1)	Tax Incentive (2)	Other special pricing (3)	None of these (4)	Don't Know (98)
Central air conditioning unit					

Exhibit 4-4
Examples of Measuring Effects Due to Other Programs (Energy Efficient Equipment)

For actions that are not typically rebated, such as weather stripping, researchers may want to determine whether these actions were taken through audit-based programs. In Exhibit 4-5, we provide an example of this type of questioning.

Question 1: (Presence of the action) Does your home have any of the following items:

	Yes (1)	No (2)	Don't Know (98)
Weather stripping or caulking around windows and/or doors			

[SKIP Questions 2-5 IF Question 1 = No]

Question 2: (Confirm the action took place during the observation period) Please indicate whether the weather stripping was purchased or installed in the past year.

	Yes (1)	No (2)	Don't Know (98)
Weather stripping or caulking around windows and/or doors			

Question 3: (Determine other program influences) Have you ever had a home energy assessment or audit, where someone from (insert utility name) or another organization came to your house and assessed your home's energy use?

1. Yes, within the past year
2. Yes, more than 1 year ago
3. No
98. Don't Know

[SKIP Question 4 and 5 IF Question 3 = No]

Question 4: (Determine other program influences) Please indicate whether your household received weather stripping or caulking around windows and/or doors installed within the past year as part of a home energy assessment.

	Yes (1)	No (2)	Don't Know (98)
Weather stripping or caulking around windows and/or doors			

Question 5: (Determine other program influences) Please indicate whether your household received weather stripping or caulking around windows and/or doors installed within the past year for free.

	Yes (1)	No (2)	Don't Know (98)
Weather stripping or caulking around windows and/or doors			

Exhibit 4-5
Examples of Low-Cost Action Likely to Have Occurred Through an Audit

As the examples demonstrate, researchers should inquire about other program influences if they are interested in determining whether the behavior was solely attributable to the feedback program. To do this efficiently, researchers should consider classifying behaviors into categories based on the measurement requirements of each. This way, researchers can ask about multiple behaviors in a single matrix (and follow up accordingly) without having to go through the sequence of questions for each individual behavior. We provide an example of how this would be conducted in Appendix A.

Measuring Self-Reported Influence without Comparison Groups

In post-only surveys without a comparison group, researchers are left to ask respondents to self-report the level of influence of a feedback program. Generally speaking, this is considered the least rigorous survey method for determining program attribution for feedback interventions. In the case of other behavioral programs that provide information only (e.g., education or information campaigns), it is generally understood that the programs are unlikely to be cited as the sole contributor to a behavior change:

"Education and information programs are generally thought of as contributing to actions; they lay the groundwork for the ability to take reasonable actions. However, they are not usually the sole reason (or even a critical reason) for taking action."[21]

While this may not be exactly the case with feedback, this is important to call out, because when asked directly whether the feedback intervention had an influence on their behaviors, respondents may underreport the influence of feedback on their actions. However, when we compare the effects of the intervention against an untreated population (i.e., a comparison group), changes are often detected.

Further, in most cases of self-reported attribution, researchers inquire about the influence of a program on a single measure. However, in the case of feedback, there are likely multiple actions that could have been influenced by feedback to varying degrees. Yet, as already discussed, it is not reasonable or feasible for researchers to ask about the intervention's effects on all actions.

To date, researchers in the energy industry have not established standard guidelines for self-reported attribution of feedback effects when it is not possible to establish comparison groups. In addition, there is growing consensus that this method of establishing attribution for feedback is unadvisable and should be avoided if possible. However, there have been methods established for other types of behavioral programs, specifically education and information programs, that may be used as a guideline to develop these types of questions. We describe one example below.

[21] Opinion Dynamics Corp., Summit Blue Consulting, Jai J. Mitchell Analytics. (2010). *PY2006-2008 Indirect Impact Evaluation of the Statewide Education & Information Programs.* Final Report: Volume I. San Francisco: Prepared for the California Public Utilities Commission. pg. 86. The report is available online at: http://www.calmac.org/publications/ODC_CPUC_0608_Edu_and_Info_Impact_Eva_VoI_Final.pdf

The Cognitive Change Index (CCI) approach has been used to measure the influence of education and information programs in California to estimate energy savings as a result of a wide range of program interventions. While energy feedback programs offer more continual and customer-specific information than education and information programs, both types of programs rely on a similar underlying concept—using information to prompt behavior change.

Behavior, particularly habitual behavior, is complex and has many influences, so rather than measuring direct program influence, the CCI measures the extent to which an information-based program caused changes in cognition (i.e., knowledge) and the number of types of actions considered. This reduces reliance on the more direct questions used in a traditional net-to-gross battery.

The CCI survey design includes three question concepts: newness of the information from the program, determination of cognitive change based on the presentation of the information from the program, and a direct influence assessment. These three concepts are then used in an algorithm to measure influence on actions taken. These types of questions may serve as a starting point for a self-reported attribution battery, but should be customized to the researchers' evaluation and/or regulatory standards. The table below demonstrates how the CCI was linked to changes in reported behaviors across multiple sectors.

Relationship to Taking Action and CCI

Cognitive Change Index	% Taking Action*			
	All	Market Actors	End User - Nonres	End User - Res
Little (1.00 - 2.50)	29%	29%	40%	17%
Some (2.51 - 5.50)	69%	71%	73%	55%
Very Much (5.51 - 7.00)	85%	87%	87%	73%

*Taking action indicated through responses to three questions

Exhibit 4-6
The Cognitive Change Index Approach[22]

Should Surveys Measure Potential Impacts on Respondent Behaviors in Both Directions?

Studies on norm-based approaches have indicated that some customers may react to some types of feedback in the opposite direction of what was intended (e.g., by increasing their consumption or changing their purchase behavior), and thus researchers should consider the potential negative effects of feedback. In particular, studies have shown that this occurs when customers are

[22] For the calculations used in developing the algorithm, see (Opinion Dynamics Corp., Summit Blue Consulting, Jai J. Mitchell Analytics, 2010). pg 88.

ideologically opposed to the energy norm,[23] or may be using less electricity than those they are benchmarked to.[24] For this reason, behaviors should be considered to have either a net-reducing or net-increasing change due to the feedback intervention. Presuming all potential effects are net-reducing may fail to measure unintended consequences of feedback.

Table 4-3 illustrates different types of behavior changes that may occur as a result of a feedback program.

Table 4-3
Examples of Purchase and Usage Behavior Changes

Type of Behavior	Energy-Reducing Influence	Energy-Increasing Influence
Purchase Behavior	Purchased sooner/earlier	Delayed purchase
	Purchased more efficient (e.g. a higher SEER rating)	Did not purchase EE or efficient behaviors or purchased a less efficient version
Usage Behaviors	Increased the frequency of the behavior	Decreased the frequency of the action
	Increased the consistency of the behavior	Decreased the consistency of the action
	Changed behavior to off-peak time of day	Changed behavior to on-peak time of day

[23] Costa, D. L., & Kahn, M. E. (2010). *Energy Conservation 'Nudges' and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment.* Cambridge, MA: National Bureau of Economic Research. Working Paper No. 15939.
[24] Carroll, E., Hatton, E., & Brown, M. (2009). *Residential Energy Use Behavior Change Pilot.* Saint Paul, MN: Presented to the Minnesota Department of Commerce Office of Energy Security.

5
CRAFTING SURVEY QUESTION TO ENHANCE ACCURACY AND REDUCE BIAS

How a survey instrument is structured has a meaningful impact on gathered information. Extensive research in the social sciences, psychology, and evaluation has shown that question wording, answer structure, question order, and other factors can radically alter what respondents say about their behavior. When focusing on feedback-related behavior, researchers have to account for both the behavior that is likely to be affected (purchase vs. behavior, as discussed above), as well as the nature of the feedback being given (its format, frequency, and content). Once these factors have been identified and delineated, an instrument can be successfully constructed.

General Rules for Effective Survey Development

While the particular focus of this report is using surveys to measure behavior change, it bears mentioning that there are several universally applicable rules to designing a good survey instrument in general. Krosnick and Presser[25] do an excellent job outlining these guidelines. The questionnaire is essentially the framework for "the conversation that takes place between researchers and respondents."[26] Conceptualizing the survey as a conversation helps identify those factors, such as leading questions, tone of voice, and word usage, that often lead to miscommunication in ordinary dialogue. Common wisdom, therefore, informs how researchers should ask questions and frame responses to give the respondent the most thorough and clear understanding of the information that the researcher wants to know. The following list outlines a standard set of general guidelines when developing a survey instrument:[27]

1. Use simple, familiar words (avoid technical terms, jargon, and slang).
2. Use simple syntax.
3. Avoid words with ambiguous meanings, i.e., aim for wording that all respondents will interpret in the same way.
4. Strive for specific and concrete wording (as opposed to general and abstract).
5. Make response options exhaustive and mutually exclusive.
6. Avoid leading or loaded questions that push respondents toward one answer.
7. Ask about one thing at a time (avoid double-barreled questions).
8. Avoid questions with double negatives.
9. Early questions should be easy and pleasant to answer, and should build a rapport between the respondent and the researcher.
10. Questions at the very beginning of a questionnaire should explicitly address the topic of the survey, as it was described to the respondent prior to the interview.

[25] Krosnick, J. A., & Presser, S. (2010). "Question and Questionnaire Design". In J. Wright, & P. Marsden (Eds.), *Handbook of Survey Research* (2nd ed., pp. 263-314). Bingley, United Kingdom: Emerald Group Publishing.
[26] Ibid, p. 2.
[27] Ibid, p 3-4.

11. Questions on the same topic should be grouped together.
12. Questions on the same topic should proceed from general to specific.
13. Questions on sensitive topics that might make respondents uncomfortable should be placed at the end of the questionnaire.
14. Filter questions that should be included, to avoid asking respondents questions that do not apply to them.

Exhibit 5-1
General Best Practice Tips

These simple points are all too often forgotten when researchers design surveys. In all survey research, researchers should follow basic survey design guidelines regardless of the topic they are studying or the mode they are using to deliver the survey.[28]

Further, while some of these points seem intuitive, researchers often over- or under-estimate respondents' knowledge of the survey subject matter. For example, engineers might think of terms like "kilowatt-hour," "compact fluorescent light bulb," "variable frequency drive," or "peak hour of usage" as easily understood, commonplace concepts. However, the public often has less knowledge of these terms than researchers expect and asking about them provides unreliable, invalid information. Even ordinary phrases like "energy efficient" and "energy conservation" can mean different things to different people without the proper context and definition.[29] Questions should also employ the active rather than passive voice, repeat nouns instead of using pronouns, and avoid possessive forms.[30] At its core, a good survey question is one that "all people answering it understand in a consistent way and in a way that is consistent with what the researcher expected it to mean."[31]

Unique Challenges of Self-Report on Behaviors

Feedback-related behaviors present a number of additional challenges for survey researchers. Asking an individual to report on their behavior is a difficult cognitive task. "A large body of research indicates that self-reports can be a highly unreliable source of data."[32] Further, different kinds of behavior present different challenges to researchers. Memorable, distinct behaviors, such as purchasing a new HVAC system, are more likely to be reliably remembered than would a habitual increase in turning off lights when leaving a room. Feedback also varies in this fashion; an energy report that comes quarterly may be more memorable than an addendum to a monthly utility bill. An in-home display (IHD) that provides information in real-time might not

[28] Survey mode is an important consideration when designing particular questions. In many cases, the presence or absence of interpersonal conversation determines the amount of candor respondents feel comfortable delivering. Telephone and in-person surveys tend to increase problems with social desirability bias and "yes-saying" bias, but are also less vulnerable to recall bias and question order effects than are self-administered mail or electronic surveys. For a more complete discussion of the effects survey mode has on response validity and reliability, please see: Bowling, A. (2005). "Mode of Questionnaire Administration Can Have Serious Effects on Data Quality". *Journal of Public Health*, 27 (3), 281-291.

[29] Schwarz, N., & Oyserman, D. (2001). "Asking Questions About Behavior: Cognition, Communication, and Questionnaire Construction". *American Journal of Evaluation*, 22 (2), 127-160.

[30] Lietz, P. (2010). "Research into Questionnaire Design: A Summary of the Literature". *International Journal of Market Research*, 52 (2), 249-272.

[31] Fowler, F. (1995). *Improving Survey Questions: Design and Evaluation*. Thousand Oaks, CA: Sage Publications.

[32] Op. Cit. (Schwarz & Oyserman, 2001, 128)

be observed on a daily basis, but may be more readily recalled than a detailed energy report. The combination of behavior and feedback details will necessarily guide the structure and timing of the survey. We expand upon the guidelines highlighted above in the following sections related to question wording, answer structure, question order, and other guidelines pertaining to feedback-induced behavior measurement.

On face value, asking respondents to report on their behaviors may seem the most efficient way of collecting data, since researchers can call fewer people and limit surveys to participants. However, self-reports can be unreliable. For a variety of psychological reasons, respondents are not good evaluators of their own behavior. By examining only feedback recipients, and only after the feedback has been delivered, researchers run the risk of consistently and seriously biasing their findings by the methods they use.

One of the problems inherent in self-reporting is the reliance on recall of events that have occurred in the past. Most respondents spend very little time thinking about their energy behavior since it is effectively unremarkable. Unlike most energy researchers, the average consumer only thinks about energy use, if at all, when paying their utility bill or when making energy-related purchases, such as lighting or HVAC. Even then, the average consumer do not have an adequate understanding of energy efficiency to provide reliable, valid estimates of their behaviors. Without the ability to accurately estimate their baseline energy-related habits, attempting to have respondents report *changes* in their behavior is especially difficult. We can classify errors in memory recall into four main categories:[33]

1. *Memories of event are not fully formed*—the less definite, concrete, or elaborate the event, the less likely the experience can be retrieved later.
2. *Errors in memories formed after the initial experience*—the level of emotional impact affects how well memories are formed after the initial experience. Often, memories are altered with inferences and learning that occurs after the experience.
3. *Retrieval failure*—where information is stored in long-term memory but the individual is unable to retrieve it. With the passing of time, similar events get grouped together to form "generic" memories where overall patterns of events can be recalled but details cannot.
4. *Reconstruction error*—where an individual tries to fill in some of the gaps in memories which are not always accurate representations of what actually happened. The individual may also "telescope" memories from one time period into another to report it to the interviewer.

While unformed memories are difficult to correct (error 1), researchers do use multiple methods to help in the memory recall process for the rest of the errors. To improve the data received from self-reporting, it is essential to ask the right questions to aid the memory retrieval process. This involves the following steps:

a. Breaking down a class of events into subclasses to help recall events separately (for example, breaking down energy saving behaviors into specific appliance-based behavior questions)—this helps rectify error 2.

b. Asking about events in a fixed chronological order—this helps rectify error 3.

[33] Op. Cit. (Lietz, 2010)

c. Listing temporal boundaries, such as major life transitions, holidays, or other events, when asking about behavior events—this helps rectify errors 3 and 4.

d. Additionally, more factual and concrete requests of information are likely to get more accurate responses and help rectify errors 3 and 4.

Besides errors in memory, there are often errors in dating of events, which could cause individuals to "telescope" events, i.e., remember events as having happened more recently than they actually did. This can lead to over-reporting of the frequency of events within a time period. This, however, cannot be fixed by simply decreasing the time period a respondent is to consider. For behaviors that are frequent, irregular, or relatively unimportant, individuals simply count the number of events having occurred in a shorter time span and then compute an answer for the longer time period. By contrast, for behaviors that are regular, individuals will already have a count and apply it to any given time period.[34] Thus, to minimize the problem of recall, surveys should be conducted as soon after the event as possible. Below, we provide a further discussion of memory recall.

Question Wording—Aided Recall

Asking respondents about their energy use behavior presents a number of challenges for researchers crafting appropriate question language. For respondents to answer behavior questions correctly, they need to have encoded the behavior in memory. Memories are encoded when they are placed in memory for future recall, because they are unique, consciously pursued actions or thoughts. Non-salient behaviors, or behaviors that are not memorable, are generally not encoded, as they often occur without the respondent actively thinking about them (such as remembering to turn off the lights after leaving a room).[35] Without an accurate memory of the behavior, respondents, often in an effort to please the interviewer, will do their best to answer these questions. This can lead to unreliable answers.

There are, however, a number of techniques that researchers should use to ameliorate this problem.

First, researchers should ask somewhat longer questions. Longer questions give the respondent more time to think about the behavior before giving their answer, and provide more context to the question. Note, this does not mean asking more complicated questions, but questions that give a clear definition of the behavior, and prompt the respondent to think about the action in question. A good rule of thumb is that no question should exceed 20 words per sentence, while containing sufficient information to elicit the respondent's comprehension.[36] Research has also shown that data quality increases if groups of questions on the same topic are preceded by a medium-length (approximately 30 words) introduction. The introduction should provide all necessary information while remaining as concise as possible.[37,38]

---

[34] Op. Cit. (Fowler, 1995)
[35] Op. Cit. (Buhr, 2011)
[36] Oppenheim, A. N. (1992). *Questionnaire Design, Interviewing and Attitude Measurement*. London: Pinter.
[37] Blair, E. E. (1977). "How to Ask Questions About Drinking and Sex: Response Effects in Measuring Consumer Behavior". *Journal of Marketing Research*, 14, 316-321.

The following series of questions are related to the energy efficient lighting products, including compact fluorescent light bulbs, or CFLs, that look like swirled tubing, resembling an ice cream cone.

In the past month, have you…

1. Purchased any CFLs for use in your home?
    - Installed any CFLs for use in your home?
2. Purchased any specialty CFLs, like flood bulbs, candelabra, dimmable, or globe bulbs to use in your home?
    - Installed any specialty CFLs in your home?

Exhibit 5-2
Standard Question Battery Introduction with Branched Follow-Up Questions

Second, the survey instrument should specify an appropriate time frame for the behavior in question. Shorter time frames are more appropriate when the behavior is non-salient. For example, a one-year timeframe is probably too long to ask about CFL purchases, but appropriate for asking about HVAC equipment installation. At the same time, researchers should be wary of using time frames that are too short. A one-month window would be appropriate for asking about the temperature chosen for washing clothes, but not for installing a ceiling fan. Feedback recall can be affected by this phenomenon as well. For example, if a utility sends monthly bill inserts, it would be more appropriate to ask about the number of bill inserts received over the last several months than asking about the past year. Another tool is the use of an alternative scenario or vignette when seeking reactions to issues that are outside the realm of the past or present, such as future energy behaviors.[39,40]

Third, the language of the question should contain cues that prompt the respondent to think about situations where the behavior is likely to have occurred. For example, a lighting researcher might be interested solely in CFL or LED purchases, but asking about the respondents' other lighting purchases, such as incandescent light bulb purchases, provides a contextual reference that may elicit more accurate responses. Further, detailed descriptions of products, behaviors, and feedback can increase the likelihood that respondents recall something accurately. For example, a monthly energy report can be described as "a piece of paper attached to your monthly bill, printed in color, which shows how much energy you have used compared to your neighbors with pie charts and tables." Here are two questions related to a typical energy conservation behavior:

---

[38] Researchers should also remember that long questions, whether read by respondents on a computer screen or by interviewers over the telephone, increase the length of surveys, which also increases the likelihood of mid-interview terminates and survey incompletes and the pros and cons of extending the length of a question should be carefully considered.

[39] Foddy, W. (1993). *Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research.* Cambridge, UK: Cambridge University Press.

[40] Measuring future intent is very difficult to accomplish with any degree of confidence, regardless of method. Most individuals are poor estimators of their past and present behavior. Asking them to generate a hypothetical scenario based on these estimates, and for that to reflect reality, is extremely cognitively difficult. While creating an approximation of future behavior may help illuminate a respondent's reaction to feedback, it is important to remember that future feedback is based on what *might* occur, not what *will* occur.

One represents a poor way of asking about turning off consumer electronics when leaving a room and one question represents a better way.

Question 1: (POOR: Unaided) How often do you turn off your computer when you're done using it?

Question 2: (GOOD: Aided) Think about the last time you finished using your computer. After you were done using it, before you left the room, did you

1. turn it off,
2. leave it on,
3. put it in stand-by mode?

Exhibit 5-3
Aided Recall and Timeframe

The first question does not encourage the respondent to envision the behavior. It places the onus of remembering the exact nature of the task on the respondent without using either the question language, or the read answer categories. It also does not specify a time period so the responses could be quite varied and difficult to analyze.

The second question places context around the behavior. It creates a mental scenario, reminding the respondent of what usually happens when they finish using the computer. It also limits the time frame to the most recent time the behavior took place. While this time might have been an isolated event, over a large sample, unique behaviors will aggregate into a reliable and valid measure of typical behavior.

Question Wording—Social Desirability Bias

In addition to recall errors, some respondents may incorrectly answer survey questions because of social desirability bias. Social desirability bias is the tendency of respondents to give answers that are more in line with the preferences of society as a whole than their actual past behaviors. For example, in medical research, survey respondents often underestimate smoking and drinking behaviors while overestimating exercise rates.[41] In social science research, political scientists have found that survey respondents overestimate their past voting behavior by nearly 15%.[42]

Evaluation surveys that ask questions about energy efficiency behaviors and attitudes may fall victim to social desirability bias errors. Social pressures, especially in jurisdictions that tend to be environmentally conscious, may encourage respondents to alter their answers to fit in with their neighbors. As such, behaviors that are not considered "appropriate," such as washing clothes in hot water or leaving lights on after exiting a room, may be underreported.

Researchers who wish to minimize social desirability bias "…should consider wording their questions to make it more socially acceptable to not take energy efficient actions."[43] This should

[41] Brace, I. (2004). *Questionnaire Design: How to Plan, Structure, and Write Survey Material for Effective Market Research.* London: Kogan Page.
[42] Holbrook, A. L., & Krosnick, J. A. (2010). "Social Desirability Bias in Voter Turnout: Reports Tests Using the Item Count Technique". *Public Opinion Quarterly*, 74 (1), 37-67.
[43] Op. Cit. (Buhr, 2011)

serve to balance the response. Researchers may be concerned about potentially letting customers feel that not taking energy efficiency action is okay, but an accurate understanding of consumer behavior is more useful for crafting effective programs and marketing campaigns than the risk that a small number of respondents will be dissuaded from conserving. Below, we present two examples of questions where social desirability bias related to energy behaviors might be present. One is an example of poor question wording, and one is an example of good question wording.

Question 1: (POOR) I am going to read a list of actions you may have taken to reduce the amount of energy your household uses. After I read each one, please tell me if this is something you regularly do.

Question 2: (GOOD) There are many people can reduce energy in their home, but we recognize that people are often too busy to take these actions. When you consider your home, do you regularly take any of these actions?

Exhibit 5-4
Reducing Social Desirability Bias

Finally, researchers can mitigate some of the problems inherent in social desirability bias by utilizing survey modes that remove interpersonal conversation (e.g., online or mail surveys). Research has found that in-person surveys increase the pressure on respondents to provide socially desirable responses.[44] For example, researchers in the UK found that respondents interviewed in person reported smoking fewer cigarettes and drinking less alcohol than respondents presented with an identical self-completed mail survey, even after accounting for sample differences.[45]

Question Wording—Other Concerns

In addition to recall error and social desirability bias, wording can also lead to reliability and validity problems if questions are phrased in ways that confuse the respondent. Among these sources of confusion are double-barreled questions,[46] negatively worded questions,[47] and variations in adverbs of frequency.[48,49] Below are examples of common energy behavior questions that highlight double-barrel, negative wording, and adverb frequency issues.

---

[44] Kreuter, F., Presser, S., & Tourangeau, R. (2008). "Social Desirability Bias in CATI, IVR, and Web Surveys". *Public Opinion Quarterly*, 72, 847-865.
[45] Tipping, S., Hope, S., Pickering, K., Erens, R., Roth, M. A., & Mindell, J. S. (2010). "The Effect of Mode and Context on Survey Results: Analysis of Data from the Health Survey for England 2006 and the Boost Survey for London". *BMC Medical Research Methodology*, 10, 84-91.
[46] Brislin, R. W. (1986). "The Wording and Translation of Research Instruments". In W. J. Lonner, & J. W. Berry (Eds.), *Field Methods in Cross-Cultural Research* (pp. 137-164). Newbury Park, CA: Sage Publications.
[47] O'Muircheartaigh, C., Krosnick, J., & Helic, A. (2000, December). *Middle Alternatives, Acquiescence, and The Quality of Questionnaire Design*, Unpublished Manuscript. Retrieved October 27, 2011, from University of Chicago Harris School of Public Policy Publications - Working Paper Series: http://harrisschool.uchicago.edu/about/publications/working-papers/pdf/wp_01_3.pdf
[48] Simpson, R. H. (1994). "The Specific Meanings of Certain Terms Indicating Differing Degrees of Frequency". *Quarterly Journal of Speech*, 30 (3), 328-330.
[49] Bradburn, N., & Miles, C. (1979). "Vague Quantifiers". *Public Opinion Quarterly*, 43 (1), 92-101.

Double-Barreled Questions

By focusing on individual appliances and lighting, researchers can better ascertain the respondent's habits, as well as how much energy their behavior is likely to save. In the example below, the "poor" examples asks about two actions, while the "good" example asks about just one.

Question 1: (POOR: Double Barreled) How often, in the past week, did you turn off the lights and television when you left the living room?

1. Always
2. Most of the time
3. About as often as not
4. Some of the time
5. Never

Question 2: (GOOD: Single Focus) How often, in the past week, did you turn off the lights when you left the living room?

1. Always
2. Most of the time
3. About as often as not
4. Some of the time
5. Never

How about turning off the television when you left the living or family room?

Exhibit 5-5
Double-Barreled Questions

Adverbs of Frequency

One of the challenges researchers face is correctly measuring how frequently a behavior takes place. In the text box below, the same question is asked, but different adverbs of frequency are used in the response categories. The meaning of these categories is highly determinative in response distribution and validity.

Question 1 (POOR: Poor Adverb Use): How often do you unplug unused cell phone chargers? Would you say…

1. Usually
2. Often
3. Sometimes
4. Rarely

Question 2 (GOOD: Better Adverb Use): How often do you unplug unused cell phone chargers? Would you say…

1. Always
2. Most of the time
3. About as often as not
4. Some of the time
5. Never

Exhibit 5-6
Adverbs of Frequency

The first question uses wording that is not universally interpreted by respondents in the same way. Usually may mean all the time for some people, or just more than half the time for others. The term "rarely," depending on the type of behavior measured, can mean only once every day, once every week, or once every year. The second question uses terms that have more standardized meanings: "always" and "never" are easily understood, and the phrase "about as often as not" presents a clear mid-point, anchoring most of the time and some of the time as intermediate categories.[50] Of course, the variation here is not so much in the question as in the answer structure given, which is our next area of focus.

Negative Wording

Though it may seem obvious, survey designers can often overlook the meaning of negative phraseology. The two questions below mean, essentially, the same thing. However, the second question is far easier for respondents to cognitively process. In question one, a respondent has to identify a list of behaviors, order it by waste (a somewhat ambiguous term) and then pick the alternative at the bottom of that list. Conversely, asking about energy savings is clearer, and the alternative of interest comes from the top of the mental list.

[50] Op. Cit. (Lietz, 2010)

Question 1 (POOR: Negative Wording): What types of energy conservation behaviors do you think are least wasteful?
(Open Ended Response)

Question 2 (GOOD: Positive Wording): What types of energy conservation behaviors save the most energy?
(Open Ended Response)

Exhibit 5-7
Negative Wording

Answer Structure—When to Use Open-Ended Questions

In most cases, open-ended questions should be limited in use; however, they do offer distinct advantages over closed-ended questions when used carefully. When crafting any survey instrument, it is important to remember that how a respondent answers a question is just as likely to be determined by the response options provided to them as the wording of the question being asked. Survey researchers use many different types of answer schemes to efficiently, and accurately, gather information from respondents. Asking questions about feedback-related behavior requires understanding how limiting, or not limiting, the answer options may bias data collection.

This consideration should not be taken lightly in the case of measuring energy behaviors. Research has shown that respondents take a greater number of actions, and more unconventional actions when treated with feedback programs. For instance, respondents may receive information that targets electricity use behavior, but as a result of the new information, take actions that are gas-specific or that affect other forms of commodity consumption, like gasoline or waste management. Essentially, respondents appear to take more actions than those suggested through feedback.[51] For this reason alone, open-ends give researchers an opportunity to understand the breadth of customer behavior without trying to cover all potential options in a single instrument.

"By allowing respondents to give an answer in their own words rather than fit their responses into a set of predefined categories, researchers can gain unexpected information. Open-ended questions also provide richer and more detailed responses than do closed-ended questions."[52] Open-ended questions can provide useful insights into behavior. For example, respondents may volunteer information about what they perceive, or misperceive, to be energy efficient behavior, adding insight into how feedback programs present savings techniques. Researchers might also learn what parts of the feedback mechanism were the most, or least, useful for respondents.

While open-ended questions present researchers with several advantages, they should be limited to a few per instrument. This is because open-ended questions are more difficult to administer and analyze. Answering an open-ended question usually requires more time on the part of the respondent (and interviewer in the case of telephone surveys). Interviewers must also be trained to get those reluctant to answer to provide responses, which must be collected verbatim. Also,

[51] Dougherty, A., & Mitchell-Jackson, J. (2009). *Ethnographic Research Findings*. San Francisco: Presented to the California Public Utilities Commission.
[52] Op. Cit. (Buhr, 2011)

generally speaking, most open-ended responses are subsequently coded by researchers for quantitative analysis. Coding takes time, requires extensive training, and is costly. Coding also may remove some of the rich detail that often accompanies open-ended responses.

Open-ended questions are also more likely than closed-ended questions to elicit a "don't know" response. Respondents either do not think of a possible behavior because they are not prompted, or they decline to speculate to avoid embarrassment. To compensate for this, open-ended questions should include wording that discourages "don't know" answers and encourages educated guessing.[53] The survey mode also affects the utility of asking open-ended questions; mail and online surveys tend to have higher rates of open-ended item response than telephone interviews, but are still more prone to missing item issues than face-to-face interviews.[54]

Another concern with open-ended questions is the degree of comparability between answers. This is especially true when recall of behaviors is the question topic. It would be difficult to confirm that two respondents who said they "usually" turned off lights when leaving the room were doing so at the same rate. Like question wording, open-ended answers that contain ambiguous terms, or are presented without context, are not as useful as more structured responses. However, open-ended questions are preferred if the question wording asks for amounts or frequency in integers, like "how many times in the past three months have you purchased CFLs?" Arbitrary categories can bias responses compared to numerical open-ended answer structures.

Further, because feedback interventions may prompt a wide range of actions, many of which the researcher may not have conceived of during the study design process, open-ended questions can serve as "catch-all" to gain more insight into the potential breadth of actions taken. However, researchers should understand the limits of this data, as noted above, and pay close attention to where such questions are placed in the instrument to prevent priming or leading the respondent prior to asking the open-ended question.

Closed-ended questions address many of the problems associated with open-ended questions. Answers are standardized, more easily collected, and are pre-coded. Further, many previously used and tested closed-ended questions already exist. These can be used as starting points for new surveys, since it is difficult to come up with exhaustive behavior lists from scratch.

There are also several behaviors that are better asked as closed-ended questions. These include frequency of behavior and simple yes/no questions on behaviors of interest, such as HVAC or weatherization purchases.

Closed-ended questions are not without their own problems, however. For a question to faithfully report on the behavior being studied, closed-ended questions have to have an exhaustive, yet concise, list of possible responses. To compensate for this difficulty, some surveys include an open-ended "other" categorical response. This is usually not effective,

---

[53] Op. Cit. (Krosnick & Presser, "Question and Questionnaire Design", 2010)
[54] Frankfort-Nachmias, C., & Nachmias, D. (1992*). Research Methods in the Social Sciences* (4th ed.). London: Edward Arnold.

however, since most respondents will avoid the "other" category and choose an explicitly stated alternative.[55]

To summarize, there are pros and cons to both open- and closed-ended questions, so in choosing which is more appropriate for a particular type of feedback-related behavior, it is essential to ask "how would I answer this question?" If a survey designer finds that they have a great deal to say about a hypothetical question, an open-ended answer structure may be preferable. Questions such as "How was the feedback most useful?", or "Have there been any major changes in your home in the past year that would have affected your electricity consumption?" all call for open ended responses. However, if a researcher finds that they can only name a few potential responses, or a scale is called for, then a closed-ended answer structure is probably the correct choice. Questions like "How frequently did you get feedback information?" and "Did you find the find the feedback information helpful?" are best asked using closed-ended options.

Answer Structure—Scales

One of the more commonly utilized closed-ended answer structures calls on respondents to choose a point on a scale that is most reflective of their opinion. Behavior scales typically focus on frequency of behavior, perceived importance of behavior, and comparisons between behaviors. How a scale is structured, labeled, and how many alternatives are offered are very important in maintaining reliable and valid measures of behavior. "Response scales are not simply 'measurement devices' that respondents use to report their behaviors. Rather, respondents may consider the range of behaviors described in the response alternatives to reflect the researcher's knowledge of or expectations about the distribution of the behavior in the 'real world.' If so, respondents may use the range of behaviors described in the response alternatives as a frame of reference in estimating and evaluating their own behavior."[56]

Scale Polarity

Scales can come in many forms, but the concepts they measure usually fall into two categories: unipolar and bipolar. A unipolar concept, such as helpfulness, varies along just one dimension. The concept is measured either on the positive or negative side of a scale, but not both. "Importance" is a concept best measured with a unipolar scale. For example, assume you are conducting a post-test only survey of program participants, and you wish to know how important feedback was in promoting smart energy decisions. The following two questions both ask about the importance of turning off unused lighting; however, the first question does not make much sense. What is the conceptual difference between somewhat and very unimportant? If the feedback is unimportant, it is not important. The second question more accurately measures the phenomenon conceptually.

---

[55] Schuman, H., & Scott, J. (1987). "Problems in the Use of Survey Questions to Measure Public Opinion". *Science*, 236, 957-959.

[56] Op. Cit. (Schwarz, Hippler, Deutsch, & and Stack, 1985, 388)

Question 1 (POOR: Bipolar): How important is turning off lights when you leave a room in order to save energy?

1. Very unimportant
2. Somewhat unimportant
3. Neither important nor unimportant
4. Somewhat important
5. Very important

Question 2 (GOOD: Unipolar): How important is turning off lights when you leave a room in order to save energy?

1. Not at all important
2. Not very important
3. Somewhat important
4. Very important

Exhibit 5-8
Scale Polarity

Conversely, a concept such as satisfaction can be measured in a bipolar fashion. Respondents can be either satisfied or dissatisfied to varying degrees.

Determining whether the concept is bipolar or unipolar, and how findings are to be described, should guide scale development. Below are a few common examples of these concepts, though this is not an exhaustive list.

Table 5-1
Common Bipolar and Unipolar Concepts

Bipolar	Unipolar
Satisfaction	Helpfulness
Agreement	Informative
Approval	Usefulness
	Importance
	Influential

Scale Length

Determining the number of scale points is also an important task. Most research indicates that 5-point and 7-point scales are the most commonly used, and are most easily rescaled for comparison.[57] The 7-point scale has been shown to be more reliable, as greater differentiation of responses increases without adding too many categories.[58] Adding more categories than eleven, however, does not increase validity, while adding collection and coding costs. In practice, shorter

[57] Dawes, J. (2008). "Do Data Characteristics Change According to the Number of Scale Points Used? An Experiment Using 5-Point, 7-Point, and 10-Point Scales". *International Journal of Market Research*, 50 (1), 61-77.
[58] Crohnbach, L. J. (1951). "Coefficient Alpha and the Internal Structure of Tests". *Psychometrica*, 16, 93-96.

scales are preferable for more concrete concepts; longer scales are more appropriate when abstract concepts are being measured.[59]

Question 1 (POOR: Scale Too Long): On a scale from one to one-hundred, with 100 being very informative, and one being not at all informative, rate the information contained in your monthly feedback form.

Question 2 (GOOD: Scale Appropriate): On a scale from one to five, with one being not at all informative, and five being very informative, rate the information contained in your monthly feedback form.

Exhibit 5-9
Scale Length

The first question undoubtedly will provide the researcher with more specific and varied responses. However, by using a scale with one hundred different values, the meanings of those values become less clear. For example, we cannot say with much certainty that a rating of 85 is truly 5 basis points less than 90, because the qualitative meanings of those numbers are probably different to two different respondents.[60] All we can gather with any certainty is that they both probably found the information quite useful. Therefore, using the shorter scale actually produces more reliable and valid results. For most behavior- and feedback-related questions, scales of ten or fewer are recommended.

Answer Labels and Branching

Labeling response categories is also a more complex task than it may first appear. Using a numeric scale that varies between zero and ten increased the tendency to use the zero category as an anchor compared to a scale that varies between one and ten.[61] In other words, respondents subjectively ascribe more weight to a zero than a one when placed at the end point of the scale, increasing its use as a location for extreme values. Adding labels to the numbers increases reliability and validity. This is doubly true when all points have labels, as opposed to just the end points.[62] Adding these labels to a five-point scale can be done fairly easily. A seven-point scale presents greater difficulty, especially if the survey is conducted by telephone. Using comprehensive labels is usually not feasible for scales with more than seven points.[63]

One way to ease this concern is to branch questions, asking a preliminary question that asks for a general characteristic or opinion (such as satisfied vs. dissatisfied), and a subsequent question that asks for refinement (very satisfied, moderately satisfied, somewhat satisfied).

---

[59] Op. Cit. (Foddy, 1993)
[60] In especially long scales, such as 1-100, respondents tend to cluster on the extremes (1, 100) the midpoint (50) and intervals of 5 and 10 rather than provide more unique ratings (such as 17 or 74). This tendency usually indicates that the scale is too long, and respondents are mentally truncating the number of rating points for cognitive efficiency.
[61] Op. Cit. (O'Muircheartaigh, Krosnick, & Helic, 2000)
[62] Krosnick, J. A., & Berent, M. K. (1993). "Comparisons of Party Identification and Policy Preferences: The Impact of Survey Format". *American Journal of Political Science*, 37, 941-964.
[63] Op. Cit. (Buhr, 2011)

Question 1 (POOR: Not Branched): Compared to turning off lights to save energy, on a scale from one to seven with one being saves a lot less energy, seven being saves a lot more energy, how much more energy do you think washing clothes in cold water saves compared to turning off the lights?

Question 2 (GOOD: Branched): Compared to turning off lights to save energy, do you think washing clothes in cold water saves:

More energy?

(If yes follow up with: Do you think it is a lot of energy, or just a little more)

About the same amount of energy?

Less energy?

(Do you think it is a lot less energy or just a little less?)

Exhibit 5-10
Answer Labels and Branching

These questions ask essentially the same thing, but the labeling and branching on the second question provides a more thorough and comprehensive scale on which the respondent can locate their belief. This provides a more accurate reflection of their true opinion.

Another consideration concerning scales is the inclusion of a middle, or neutral, category. Many studies have shown that more respondents will choose the middle position if it is offered than if they have to volunteer that response.[64] At the same time, researchers are split on whether removing the middle category systematically affects responses. Researchers should look to the concept they are attempting to measure before constructing the scale. If there is a genuine middle or neutral category to the subject matter, then the instrument should include one.

Additionally, scale responses should be designed to account for primacy and recency effects. The primacy effect is the observed tendency for respondents to pick early alternatives in a scale over later alternatives. The recency effect occurs when respondents pick the later alternatives more frequently. The recency effect is more commonly observed when lists are read over the phone. While the phenomenon is more common on extremely long lists (16 or more alternatives), it is often advisable to randomize the reading of scales, so as to minimize the potential impact of either primacy or recency effects.

As with scale length, researchers have to determine if a "don't know" response should be included, and if it should be read on a phone survey. For phone surveys, "don't knows" are typically not read.[65] When using mail or web surveys, researchers are advised to include "don't know" responses to minimize the risk that participants will terminate because they feel they have no opinion or information on a particular question.

[64] Krosnick, J. A., & Schuman, H. (1988). "Attitude Intensity, Importance, and Certainty and Susceptibility to Response Effects". *Journal of Personality and Social Psychology*, 54, 940-952.
[65] Op. Cit. (Buhr, 2011)

Question Order

In addition to question wording and answer structure, the order of questions in a survey has been shown to affect how respondents answer questions. Three major types of problems associated with question order effects have been shown in research; these include "part-whole," "part-part," and saliency effects.

Part-whole combination effects arise when a concept, such as energy usage, is divided into general and specific questions. Problems occur if general questions are asked after specific questions. In theory, asking a general question after a specific question appears to remove the specific concern from the general question decision-calculus. For example, asking a person about the relative importance of conserving energy should proceed from a general question, such as "how important is it to save energy?" to the specific "how important is it to purchase energy efficient light bulbs to save energy?" Reversing the order would imply that the general question does not relate to light bulb purchases, as they have already been asked about, which alters the final question's meaning to the respondent. [66]

Broad questions on behavior:
"Since the feedback, have you changed the way you consume energy?"
Narrower, task or purchase oriented behavior questions:
"Have you purchased any CFLs since the feedback?"
Secondary task or purchase-oriented behavior questions:
"Are all of these CFLs ENERGY STAR-certified products?"

Exhibit 5-11
Question Order

"Part-part" question order effects occur when respondents answer questions at the same level of specificity, but attempt to "normalize" answers to subsequent questions to match opinions related to the first question. A classic example of this was found in a survey where respondents were asked if the Soviet Union should admit journalists from the United States, and then asked if the U.S. should allow Soviet journalists to enter the country. Those who were asked these questions in reverse order were far less likely to support the U.S. admitting Soviet reporters than were those who were asked about U.S. reporters being allowed in the U.S.S.R. first.[67] In evaluation, asking about a popular program before a less popular program may prompt respondents to rate the second program more favorably than if they were asked about it first.

Salience-related question order effects occur when a topic is raised earlier in the survey, resulting in respondents who are more inclined to answer subsequent questions on the topic. This is especially true if the topic is at all controversial. For example, asking if "climate change is real" early in a survey might prompt respondents to answer subsequent open-ended questions about problems facing the nation with "climate change." To minimize this effect, researchers should ask intervening questions between possible prompts and subsequent related questions, and keep language as neutral as possible.

---

[66] Bates, N., Martin, E., DeMaio, T., & de la Puente, M. (2006). *Questionnaire Effects on Measurements of Race and Spanish Origin. Research Report Series.* Washington, DC: Statistical Research Division, US Census Bureau.
[67] Op. Cit. (Lietz, 2010)

Accurately answering questions about behavior is actually a difficult cognitive task, and the way in which questions are posed dramatically affect the answers obtained.[68] To address this challenge, this section has provided a set of guidelines to aid researchers in developing survey instruments to tease out the specific behaviors that feedback interventions may cause. For convenience, a quick reference guideline checklist can be found in Appendix C.

[68] Op. Cit. (Schwarz & Oyserman, 2001)

6
CONCLUSIONS

While studies have shown that feedback programs can generate energy savings, it is important to understand the underlying sources of the energy savings—which means understanding and studying the various behaviors exhibited by participants. Given the nature of feedback information and the limitations of extensive monitoring to obtain end-use data, researchers can leverage the guidelines discussed in this report to help their survey efforts meet the greatest levels of rigor within budgetary and timing limitations. Below is a summary/checklist of the important questions to ask to ensure a responsible and defensible study.

First, researchers must understand how the survey effort fits into the **Overall Study Design**. This can be done by answering the following questions:

- Is this research meant to complement an impact evaluation? If so, how?
- Are there any requirements of other efforts to which this study must conform? The use of a control group? Sampling?

Researchers must determine the **Overall Goals of the Survey Research**. This can be done by answering the following questions:

- What are the unique goals of the survey research? What is the intended use of the findings?
  - Is the survey being used to estimate attribution or energy savings? If so, what is the greatest level of objectivity and rigor that the study can support?
  - Is the goal to inform a process evaluation? If so, then lower levels of rigor may be acceptable.
  - Is there a need to account for double counting/"other program participation"? Is double counting or other program participation being accounted for elsewhere in the overall study?

Researchers must determine and choose an appropriate **Survey Study Design**. This can be done by answering the following questions:

- Has the program launched? If not, is a pre-test feasible?
- Can researchers set up a comparison group within the time and budget available?
- What is the required level of rigor in attributing measured behavior change to the feedback intervention?

Researchers must define to whom the results will apply and thus **Sample** appropriately by answering the following question:

- Is the study being used to study the general population or a specific subgroup?
- What level of precision is required for each measured behavior? How will this vary based on the expected incidence of the behavior in the population?

Researchers must understand the various **Survey Modes** available for contacting individuals and must decide the most appropriate one to use for the research study. This can be done by answering the following questions:

- How were feedback participants contacted and what data were available (e.g., online, by phone)?
- If a comparison group is being used, is there comparable contact information for the comparison group?

Researchers must determine the range of behaviors being targeted and the focus of the survey efforts so that the **Focus and Length of Survey** is appropriate**.** This can be done by answering the following questions:

- Is the survey capturing the types of behaviors that are expected to change during the observation period?
- Have the behaviors been carefully prioritized?
- Are there certain behaviors that require greater level of precision than others?
- Is the survey structured to move from general to specific questions?

Researchers must **Minimize Bias and Error** within the survey instrument. This can be done by answering the following questions:

- Is the survey designed appropriately to remove potential biases (such as leading questions, tone of voice, and word usage)?
- Is the survey designed appropriately to obtain quality data? Has the researcher used the checklist in Appendix C prior to fielding the survey instrument?

As demonstrated by this report, the specific guidelines for survey design and implementation are largely dependent on the answers to these questions. If researchers carefully consider each of these questions through their survey research, they will be well positioned to deliver rigorous and defensible findings within the specific goals and constraints of the study.

A
APPENDIX: REFERENCES

Bates, N., Martin, E., DeMaio, T., & de la Puente, M. (2006). *Questionnaire Effects on Measurements of Race and Spanish Origin. Research Report Series.* Washington, DC: Statistical Research Division, US Census Bureau.

Blair, E. E. (1977). “How to Ask Questions About Drinking and Sex: Response Effects in Measuring Consumer Behavior". *Journal of Marketing Research , 14*, 316-321.

Bowling, A. (2005). "Mode of Questionnaire Administration Can Have Serious Effects on Data Quality". *Journal of Public Health , 27* (3), 281-291.

Brace, I. (2004). *Questionnaire Design: How to Plan, Structure, and Write Survey Material for Effective Market Research.* London: Kogan Page.

Bradburn, N., & Miles, C. (1979). “Vague Quantifiers”. *Public Opinion Quarterly , 43* (1), 92-101.

Brislin, R. W. (1986). “The Wording and Translation of Research Instruments”. In W. J. Lonner, & J. W. Berry (Eds.), *Field Methods in Cross-Cultural Research* (pp. 137-164). Newbury Park, CA: Sage Publications.

Buhr, T. (2011). "Think Before You Do: The Importance of Survey Design in Program Evaluation". *International Energy Program Evaluation Conference.* Boston, MA.

Carroll, E., Hatton, E., & Brown, M. (2009). *Residential Energy Use Behavior Change Pilot.* Saint Paul, MN: Presented to the Minnesota Department of Commerce Office of Energy Security.

Costa, D. L., & Kahn, M. E. (2010). *Energy Conservation 'Nudges' and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment.* Cambridge, MA: National Bureau of Economic Research. Working Paper No. 15939.

Crohnbach, L. J. (1951). "Coefficient Alpha and the Internal Structure of Tests”. *Psychometrica , 16*, 93-96.

Dawes, J. (2008). “Do Data Characteristics Change According to the Number of Scale Points Used? An Experiment Using 5-Point, 7-Point, and 10-Point Scales”. *International Journal of Market Research , 50* (1), 61-77.

Dillman, D. (2007). *Main and Internet Surveys – The Tailored Design Method* (2nd ed.). New Jersey: John Wiley and Sons Inc.

Dougherty, A. (2011). *Massachusetts Cross-Cutting Behavioral Program Evaluation.* Oakland, CA: Presented to the Massachusetts Energy Efficiency Advisory Council.

Dougherty, A., & Mitchell-Jackson, J. (2009). *Ethnographic Research Findings.* San Francisco: Presented to the California Public Utilities Commission.

EPRI. (2010). *Electricity-Use Feedback in the Commercial Sector: Examining the Potential for Building Occupant Behavior Change.* Palo Alto, CA: EPRI. Report No. 1021193.

EPRI. (2010). *Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols.* Palo Alto, CA: EPRI. Report No. 1020855.

EPRI. (2011, Forthcoming). *Understanding Electric Utility Customers: What We Know and What We Need to Know (DRAFT).* Palo Alto, CA: EPRI. Report No. 1023562.

Foddy, W. (1993). *Constructing Questions for Interviews and Questionnaires: Theory and Practice in Social Research.* Cambridge, UK: Cambridge University Press.

Fowler, F. (1995). *Improving Survey Questions: Design and Evaluation.* Thousand Oaks, CA: Sage Publications.

Frankfort-Nachmias, C., & Nachmias, D. (1992). *Research Methods in the Social Sciences* (4th ed.). London: Edward Arnold.

Hall, N., De Cotis, P., Kushler, M., Megdal, L., & Vine, E. (2007). *An Evaluation Approach for Assessing Program Performance from the State Energy Program.* Washington, DC: Presented to the Office of Weatherization and Intergovernmental Programs within the U.S. Department of Energy.

Holbrook, A. L., & Krosnick, J. A. (2010). “Social Desirability Bias in Voter Turnout: Reports Tests Using the Item Count Technique". *Public Opinion Quarterly , 74* (1), 37-67.

Kreuter, F., Presser, S., & Tourangeau, R. (2008). “Social Desirability Bias in CATI, IVR, and Web Surveys”. *Public Opinion Quarterly , 72*, 847-865.

Krosnick, J. A., & Berent, M. K. (1993). “Comparisons of Party Identification and Policy Preferences: The Impact of Survey Format”. *American Journal of Political Science , 37*, 941-964.

Krosnick, J. A., & Presser, S. (2010). "Question and Questionnaire Design". In J. Wright, & P. Marsden (Eds.), *Handbook of Survey Research* (2nd ed., pp. 263-314). Bingley, United Kingdom: Emerald Group Publishing.

Krosnick, J. A., & Schuman, H. (1988). “Attitude Intensity, Importance, and Certainty and Susceptibility to Response Effects”. *Journal of Personality and Social Psychology , 54*, 940-952.

Lietz, P. (2010). "Research into Questionnaire Design: A Summary of the Literature". *International Journal of Market Research , 52* (2), 249-272.

Lutzenhiser, L. (2002). *An Exploratory Analysis of Residential Electricity Conservation Survey and Billing Data.* Pullman, WA: Presented to the California Public Utilities Commission.

Megdal, L. M., Pertusiello, S., & Jacobson, B. (1996). *All These Years Measuring Free Ridership and Now We Measure a Portion of These as Caused by Market Transformation.* Arizona: Presented to the Association of Energy Services Professionals.

Megdal, L. M., Pertusiello, S., & Jacobson, B. (1997). Measuring Market Transformation Due to Prior Utility Efforts. *Energy Evaluation Conference.* Chicago.

O'Muircheartaigh, C., Krosnick, J., & Helic, A. (2000, December). *Middle Alternatives, Acquiescence, and The Quality of Questionnaire Design*, Unpublished Manuscript. Retrieved October 27, 2011, from University of Chicago Harris School of Public Policy Publications - Working Paper Series: http://harrisschool.uchicago.edu/about/publications/working-papers/pdf/wp_01_3.pdf

Opinion Dynamics Corp., KVDR Inc, and Jai J. Mitchell Analytics. (2010). *PY2006-2008 Indirect Impact Evaluation of the Statewide Marketing and Outreach Programs. Final Report: Volume I.* San Francisco: Prepared for the California Public Utilities Commission.

Opinion Dynamics Corp., Summit Blue Consulting, Jai J. Mitchell Analytics. (2010). *PY2006-2008 Indirect Impact Evaluation of the Statewide Education & Information Programs. Final Report: Volume I.* San Francisco: Prepared for the California Public Utilities Commission.

Oppenheim, A. N. (1992). *Questionnaire Design, Interviewing and Attitude Measurement.* London: Pinter.

Owens, L. K. (2005). *Introduction to Survey Research Design.* Chicago: University of Illinois at Chicago Survey Research Laboratory.

Randazzo, K. V., & Peters, J. S. (2011). *Reconsidering What We Measure: Residential Decision-Making and Proposed Standard Questionnaire Items.* San Diego, CA: Presented to Southern California Edison and Pacific Gas & Electric.

Ridge, R., Willems, P., Itron, J. F., & Randazzo, K. V. (2009). "The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating the Net-to-Gross Ratio". *International Energy Program Evaluation Conference.* Portland.

Roberts, C. (2007). *Mixing modes of data collection in surveys: A methodological review.* London: Briefing Paper for The Economic and Social Research Council, National Centre for Research Methods.

Schuman, H., & Scott, J. (1987). "Problems in the Use of Survey Questions to Measure Public Opinion". *Science , 236*, 957-959.

Schwarz, N., & Oyserman, D. (2001). "Asking Questions About Behavior: Cognition, Communication, and Questionnaire Construction". *American Journal of Evaluation , 22* (2), 127-160.

Schwarz, N., Hippler, H., Deutsch, B., & and Stack, F. (1985). "Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments". *Public Opinion Quarterly , 49*, 388-395.

Simpson, R. H. (1994). "The Specific Meanings of Certain Terms Indicating Differing Degrees of Frequency". *Quarterly Journal of Speech , 30* (3), 328-330.

Tipping, S., Hope, S., Pickering, K., Erens, R., Roth, M. A., & Mindell, J. S. (2010). "The Effect of Mode and Context on Survey Results: Analysis of Data from the Health Survey for England 2006 and the Boost Survey for London.". *BMC Medical Research Methodology , 10*, 84-91.

B
APPENDIX: SURVEY INSTRUMENT EXAMPLE

A Note on this Example

This survey offers an example of how to construct a survey instrument that collects necessary information in the provided in the examples in Sections 4 and 5. Here, we incorporate the core data collection requirements of Section 4 while also applying some of the guidelines of Section 5. Note, this example focuses on collecting data on specific energy behaviors, and does not provide examples for process and evaluative questions.

This survey instrument should be customized to the survey design and the feedback intervention. This instrument is meant to serve as an example, not a turn-key instrument for all feedback efforts.

It is important to call out that constructing an instrument to measure multiple behaviors requires trade-offs in question wording and design—when measuring only a select few behaviors, survey wording guidelines in Section 5 can be followed precisely. When measuring more behaviors, researchers must use the guidelines as much as possible within the limitations of the instrument (budget and length). Where we stray from our guidelines in Section 5, we call out why this decision was made in a footnote.

Survey Background

Here, we provide an example of a multi-mode survey (both internet and phone) for a treatment and comparison group pre-post survey for an internet-based feedback program. Question phrasing is necessarily different between modes. For all questions, there is an internet and phone script, signified by "*INT*" and "*PHN*."

This instrument can be can be used for any form of feedback program; the primary goal is to collect information on energy-specific behaviors overall and, for this reason, it is not feedback-specific and does not go into great detail on any one behavior. In this example, we provide a range of behaviors that can be included, however the specific behaviors included should be determined as part of the research design process.

For this instrument, both treatment and comparison groups are surveyed prior to program treatment and after one year of program exposure (see Section 3 for more information on this survey design method). As described in Section 3, to determine the effects of the feedback intervention, the researchers will analyze the difference between the pre- and post- results of the participant and comparison group at t_1 and t_2. For this reason, the survey does not include any direct questions about the effects of the feedback intervention.

Survey Example

Screening Questions

Screener questions are used to select the targeted respondent for the survey. Please note that screeners should be carefully considered to determine whether or not they introduce bias. Typical screening criteria include the following:

- Recall receiving feedback
- Speaking to the person responsible for paying the utility bills
- Confirm whether employee of a utility company
- Confirm address/primary residence
- Demographics

Q1. **(Respondent has the ability to take action)** Do you or members of your household own this home or do you rent?

1. Own/Buying
2. Rent/Lease
98. (Don't Know)*
99. (Refused)

*Note for internet or phone surveys, don't know and refused options are not shown or read out to respondents. This is reflected throughout the following survey questions.

Purchase Behaviors

Q2**. (Confirm presence of the behavior)**

INT: I am going to list items or equipment that might be in your home.

Does your home have any of the following items or equipment? [ROTATE; check boxes]

PHN: I am going to list items or equipment that might be in your home.

Does your home have a… [ROTATE; 1=Yes, 2=No, 98= (Don't Know), 99= (Refused)]

	Yes (1)	No (2)	Don't Know (98)
a. [SKIP IF Q1 = 2] Central air conditioning unit			
b. Room or wall air conditioning unit			
c. [SKIP IF Q1 = 2] Clothes washing machine[69]			
d. [SKIP IF Q1 = 2] Clothes dryer			
e. [SKIP IF Q1 = 2] Dishwasher			
f. Television			
g. Printer			

[69] Note this and other water-heating actions will be analyzed only for those customers who have electric water heating.

h. Computer			
i. Weather stripping or caulking around windows and/or doors			
j. Energy Star light bulbs or compact fluorescent lights, also known as CFLs[70]			
k. Insulated outlets and/or light switches			
l. Fluorescent lights that are not compact (e.g. fluorescent bulbs that are longer and thinner than CFL)			
m. Motion sensors (e.g., for lighting)			

[ASK IF Q2c=1]

Q3. **(Determine level of energy savings)** Is your washing machine front-load or top-load?[71]

1. Front-load
2. Top-load
98. (Don't Know)
99. (Refused)

[ASK if any Q2a-m=1]

Q4. **(Confirm the behavior occurred in the observation period)** Did you purchase or install any of the items or equipment we just discussed in the last year?[72]

1. (Yes)
2. (No)
98. (Don't Know)
99. (Refused)

[ASK IF Q4=1]

Q5. *INT:* Please indicate which of the following items or equipment your household purchased or installed in the past year. [check boxes]

PHN: Did your household purchase or install [INSERT EACH Q2=1] in the past year? [1=Yes, 2=No, 98= (Don't Know), 99= (Refused)]

[70] Because this survey is quantifying multiple types of behaviors, we chose to inquire about the presence of CFLs (yes/no) rather than the number installed. If this is a metric of interest (for example a promoted action), researchers may want to consider a follow-up question to establish quantity.

[71] Note that certain equipment, such as a washing machine, has different criteria for determining the level of efficiency than simply asking about ENERGYSTAR. Questions like this should be included as relevant to the behavior of interest.

[72] By asking this question first, rather than leading with questions on efficiency, we reduce bias due to social desirability while also determining whether the feedback intervention could have impacted the action for post surveys.

	Yes (1)	No (2)	Don't Know (98)
a. [INSERT EACH Q2a-m=1]			
b.			

Q6. (Determine the level of energy savings)

INT: Please indicate whether the items or equipment you installed in the last year were additional units or replacement units (i.e., replaced an older model) [check boxes]

PHN: Was the [INSERT EACH Q2a-m=1] you installed in the last year an additional unit or replacement of existing [INSERT EACH Q2a-m=1]? [1=Additional, 2=Replacement, 98= (Don't Know), 99= (Refused)]

	Additional Unit/ System/Product (1)	Additional Unit/ System/Product (2)	Don't Know (98)
a. [INSERT EACH Q2a-m=1]			
b.			

Q7. **(Confirm level of energy savings)**

INT: Are any of the items or equipment you purchased in the last year ENERGY STAR qualified [for the internet version, and ENERGY STAR logo will be shown as a point of reference for the respondent] [check boxes]

PHN: To the best of your knowledge, is/are the [INSERT EACH Q2a-m=1, DO NOT INCLUDE Q2j] you purchased in the last year ENERGY STAR qualified? ENERGY STAR qualified products have a small, square, blue logo with the "ENERGY STAR" in white letters. [1=Yes, 2=No, 98= (Don't Know), 99= (Refused)]

	Yes[73] (1)	No (2)	Don't Know (98)
a. [INSERT EACH Q2a-h=1][74]			
b.			

[ASK IF Q4=1]

[73] Images, such as the ENERGY STAR logo, can be used to help establish efficiency levels. For phone surveys, interviewers should use a description to help prompt recall.

[74] Note that we do not ask about the level of efficiency for those actions that are inherently efficient, such as weather stripping.

[ASK FOR EACH Q2a-h=1][75]

Q8. **(Determine other program influences)**

INT: Did you receive a rebate on the cost of any of the items or equipment you installed or purchased in the last year, or a tax incentive for any of these items? [check boxes]

PHN: Did you receive a rebate or tax incentive for the…? [INSERT EACH Q2a-h=1] [1=Rebate, 2=Tax Incentive, 3= (Other special pricing mentioned), 4= (None of these / no special pricing), 98= (Don't know), 99= (Refused)]

	Rebate (1)	Tax Incentive (2)	Other special pricing (3)	None of these (4)	Don't Know (98)
a. [INSERT EACH Q2a-h=1]					
b.					

[ASK IF Q4=1]

Q9. **(Determine other program influences)** Have you ever had a home energy assessment or audit, where someone from (insert utility name) or another organization came to your house and assessed your home's energy use?

1. Yes, within the past year
2. Yes, more than 1 year ago
3. No
98. (Don't Know)

[ASK IF Q9=1]

Q10. *INT:* Please indicate whether your household received any of the items you installed within the past year through the audit? If so, please tell us how or where you received the items. [check boxes]

PHN: Did you receive [INSERT EACH Q2i-m=1] through the audit? [1=Yes, 2=No, 98= (Don't Know), 99= (Refused)]

	Yes (1)	No (2)	Don't Know (98)	*INT* Q11. If Yes: Specify source of item
a. [INSERT EACH Q2i-m=1][76]				[OPEN END]
b.				

[75] Note we ask about other potential influences on the efficient purchase, but only for those items that may qualify for a rebate or tax incentive.

[76] Note we only ask about the influence of an audit for those actions that may have been directly impacted by the audit through a direct install. Researchers may want to consider whether audit-feedback crossover matters for other actions, such as appliances, as both may have contributed to the respondent's decision.

Usage Behaviors

Q11. **(Presence of the action)**

INT: There are many people can reduce energy in their home, but we recognize that people are often too busy to take these actions. When you consider your home, do you regularly[77] take any of these actions?[78] [ROTATE; check boxes]

PHN: There are many people can reduce energy in their home, but we recognize that people are often too busy to take these actions. When you consider your home, do you regularly take any of these actions? … [ROTATE; 1=Yes, 2=No, 98= (Don't Know), 99= (Refused)]

	Yes (1)	No (2)	Don't Know (98)
a. Hang laundry to dry			
b. Wash laundry in cold water			
c. [ASK IF Q1c=1] Fully load washing machine			
d. [ASK IF Q1e=1] Fully load dishwasher			
e. Turn off lights in unoccupied rooms			
f. [ASK IF Q1h=1] Turn off computers at night/when not in use			
g. [ASK IF Q1h=1] Put computer(s) to sleep			
h. [ASK IF Q1f=1] Turn off TV(s) when not in use			
i. Switch off power strips or unplug devices when not in use (chargers, TVs, stereos, etc)			
j. Take short showers			
k. Use a portable window fan			
l. Maintain your heating and cooling system			
m. Reduce the water heater temperature			
n. Clean or replace air filters			
o. Clear the area around vents			
p. Make sure refrigerator seals are tight			
q. Clean refrigerator coils			
r. Check refrigerator temperature			

[ASK IF ANY Q11=1]

Q12. **(Confirm the action took place during the observation period)** Did you start taking any of the actions we just discussed in the past year?

[77] We use the term "regularly" because we are asking about multiple actions at one time. If the instrument were designed to ask each behavior individually, behavior-specific frequencies would be developed to establish absolute frequencies (such as "every night") instead of the term "regularly," which is subject to interpretation.

[78] This question phrasing aims to reduce social desirability bias. Similar wording should be considered when asking directly about actions. In our equipment example, we led up to determining efficiency levels, mitigating bias along the way.

1. (Yes)
2. (No)
98. (Don't Know)
99. (Refused)

[ASK IF Q12=1] (Confirm the action took place during the observation period)

Q13. *INT:* Did you start taking any of these actions within the past year? [check boxes]

PHN: Did you start [INSERT EACH Q11a-r=1] within the past year? [1=Yes, 2=No, 98= (Don't Know), 99= (Refused)]

	Yes (1)	No (2)	Don't Know (98)
a. [INSERT EACH Q11a-r=1]			
b.			

Q15. Are there any other actions you started taking in the past year to save energy in your home, besides the actions you've already mentioned? Please list up to three actions. [MULTIPLE RESPONSE]

a. [OPEN RESPONSE]

b. [OPEN RESPONSE]

c. [OPEN RESPONSE]

d. No actions

The next set of questions is about the actions you've been taking for more than a year.[79]

[ASK IF any Q12a-r=1]

Q16. **(Determine level of efficiency of the action)**

INT: Please indicate whether you increased or decreased the frequency with which you took these actions in the past year compared with previous years. [check boxes]

PHN: Did you [INSERT each Q12a-r=1] more or less frequently in the past year compared with previous years? [1=Increased Frequency, 2=Decreased Frequency, 3=No change in frequency, 98= (Don't Know), 99= (Refused)]

	Increased Frequency (1)	Decreased Frequency (2)	No change in frequency (3)	Don't Know (98)
a. [INSERT EACH Q12a-r=1]				
b.				

[79] We recognize that usage behaviors can be affected in multiple ways. Here, if respondents indicate that their behavior is not new, we give them the opportunity to report on changes in frequency (increase, decrease, stayed the same). Similar to our previous discussion on the term "regularly," if we were asking about each behavior at a time, the frequency anchors would be behavior-specific.

Demographic Questions

Following are the questions recommended by the DOE*:

- Type of Dwelling
- Presence and number of room air conditioners (presence of air conditioner question asked within equipment ownership)
- Presence and programming of a programmable thermostat
- Number of people living in the home (by age)
- Presence of person with chronic illness or medical disability in the home
- Presence of person in the home on Monday to Friday between 1 PM and 5 PM
- Presence of person working full time and if person works out of the home
- Remembrance of receiving the pilot solicitation
- Usefulness of solicitation information
- Primary language spoken in home
- Level of education
- Income range

Additional questions can be asked based on utility preferences:

- Gender
- Race
- Age
- Housing Characteristics (type of residence, square footage of home, year of construction, number of rooms/bathrooms)

*U.S. Department of Energy Smart Grid Investment Grant. Technical Advisory Group Guidance Document #9. Topic: Preferences for DOE Required Data Collection via Survey Instruments (July 6, 2011)

C
SURVEY REVIEW CHECKLIST

The survey review checklist provided here offers researchers a tool for systematically ensuring that the survey instrument aligns with the guidelines outlined in Section 5.

Name of Survey: _____ Date of Review: _____

Description of the Data Collection Modes			
☐ Mail ☐ Telephone ☐ Web-based ☐ Multi-mode			
Question Review			
Satisfied	Not Satisfied	If "Not Satisfied" - indicate question number	Item
			1. Use simple, familiar words (avoid technical terms, jargon, and slang)
			2. Use simple syntax
			3. Avoid words with ambiguous meanings, i.e., aim for wording that all respondents will interpret in the same way
			4. Strive for wording that is specific and concrete (as opposed to general and abstract)
			5. Make response options exhaustive and mutually exclusive
			6. Avoid leading or loaded questions that push respondents toward one answer
			7. Ask about one thing at a time (avoid double-barreled questions)
			8. Avoid questions with double negatives
			9. Early questions should be easy and pleasant to answer, and should build a rapport between the respondent and the researcher
			10. Questions at the very beginning of a questionnaire should explicitly address the topic of the survey, as it was described to the respondent prior to the interview
			11. Questions on the same topic should be grouped together
			12. Questions on the same topic should proceed from general to specific
			13. Questions on sensitive topics that might make respondents uncomfortable should be placed at the end of the questionnaire
			14. Filter questions that should be included, to avoid asking respondents questions that do not apply to them
			15. Appropriate time frames are used for behavior questions
			16. Screening questions present when needed
			17. Question stem and responses agree in terms of concept and scales
			18. Don't know, refusal, not applicable used consistently throughout document
			19. All scales similar in terms of points and modality. If not, it is clearly indicated in question that the scale has changed
			20. Probes are clearly stated for depth guides
			21. Unusual concepts have introductory statement
			22. It is clear what the interviewer should and should not read for questions and responses

Export Control Restrictions

Access to and use of EPRI Intellectual Property is granted with the specific understanding and requirement that responsibility for ensuring full compliance with all applicable U.S. and foreign export laws and regulations is being undertaken by you and your company. This includes an obligation to ensure that any individual receiving access hereunder who is not a U.S. citizen or permanent U.S. resident is permitted access under applicable U.S. and foreign export laws and regulations. In the event you are uncertain whether you or your company may lawfully obtain access to this EPRI Intellectual Property, you acknowledge that it is your obligation to consult with your company's legal counsel to determine whether this access is lawful. Although EPRI may make available on a case-by-case basis an informal assessment of the applicable U.S. export classification for specific EPRI Intellectual Property, you and your company acknowledge that this assessment is solely for informational purposes and not for reliance purposes. You and your company acknowledge that it is still the obligation of you and your company to make your own assessment of the applicable U.S. export classification and ensure compliance accordingly. You and your company understand and acknowledge your obligations to make a prompt report to EPRI and the appropriate authorities regarding any access to or use of EPRI Intellectual Property hereunder that may be in violation of applicable U.S. or foreign export laws or regulations.

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

1021961

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