EM&V Uniform Methods Project: Survey Design and Implementation Protocols for Estimating Gross Savings

Robert Baumgartner, Tetra Tech

Survey research plays an important role in evaluation, measurement, and verification (EM&V) methods for energy-efficiency program evaluations, as the majority of energy-efficiency program evaluations use survey data.

EM&V efforts are only as accurate as the data used in analyses. However, despite the prominent role of survey research in EM&V for energy-efficiency programs, it is rare to see descriptions of survey research methods and procedures presented in sufficient detail for readers to evaluate the quality of data used in generating the findings.

This chapter presents an overview of best practices for designing and executing survey research to estimate gross energy savings in energy-efficiency evaluations. A detailed description of the specific techniques and strategies for designing questions, implementing a survey, and analyzing and reporting the survey procedures and results is beyond the scope of this chapter. So for each topic covered below, readers are encouraged to consult articles and books cited in the References and Resources section, as well as other sources that cover the specific topics in greater depth.

1 Introduction

This chapter focuses on the use of survey methods to collect data for estimating gross savings from energy-efficiency programs. Thus, this section primarily addresses survey methods used to collect data on the following:

- Characteristics of energy consumers (residential and nonresidential), include appliance and equipment ownership and reported behaviors. The results of a well-designed survey help in estimating gross savings attributable to energy-efficiency programs.
- Verification of installation, hours of use, operating conditions, and persistence of new energy-efficient equipment.
- Estimation of self-reported changes in behaviors used by households or businesses in response to energy feedback information.
- Market characteristics and sales of appliances and equipment. This information is used to establish a baseline for evaluating the impact of energy-efficiency programs on market transformation.
- Estimation of the response to retrofit and energy audit programs designed to increase the efficiency of energy use in households and businesses.

As surveys also provide the primary means of identifying and assessing non-programmatic effects, such as freeridership, spillover, and market effects, they provide the basis for calculating net savings. (These applications are described in a separate chapter.)

In defining and describing best practices for survey research, the American Statistical Association states: "The quality of a survey is best judged not by its size, scope, or prominence, but by how much attention is given to dealing with the many important problems that can arise."[1] Evaluating survey research and survey data in the manner described in that quotation requires:

- An understanding of the different sources and problems that can arise in designing and executing survey research; and
- An awareness of best practices for preventing, measuring, and dealing with these potential problems.

This chapter contains guidelines for selecting appropriate survey designs and recommends some administration procedures for different types of energy-efficiency EM&V surveys.

2 The Total Survey Error Framework

Total Survey Error (TSE) is a framework that allows researchers to make informed decisions for maximizing data quality by minimizing total survey error within the constraints of a given research budget (Groves, 2010). The TSE framework (widely used as a paradigm in survey research) is applied in evaluating specific types of survey research design. It is also used in evaluating the survey data collected to measure the behaviors of energy consumers for estimating gross savings resulting from energy-efficiency programs.

In addition to TSE, other sources of error—such as modeling decisions, low internal and/or external validity, and use of an inappropriate baseline—may also be present in estimates of gross energy savings. However, this chapter deals only with TSE. (Other chapters discuss the appropriate use of modeling and research design for specific end-uses, such as lighting, HVAC, and retrofits.)

For this chapter, the following key terms require definition:

- ***Population of Interest.*** The population to which results are to be generalized, sometimes known as the "target" population.
- ***Sampling frame.*** A directory, database, or list covering all members (or as many as possible) of the population of interest.
- ***Sampling element and unit of analysis.*** Persons, groups, or organizations from which data are to be collected.
- ***Survey errors.*** Deviation of a survey response from its underlying true value, caused by random sampling error, coverage error, nonresponse error, and measurement error.
- ***Mode-effects.*** Differences in the same measure, arising from differences in the mode of data collection used (such as interviewer-administered and self-administered surveys).

[1] American Statistical Association, 1980.

2.1 TSE Framework for Evaluating Survey and Data Quality

TSE provides a basis for developing a cost-benefit framework by describing statistical properties (or fitness for use) of survey estimates that incorporate a range of different error sources. The development of a cost-benefit framework is beyond the scope of this chapter; however, Groves (1989) describes how to reduce errors using the principles of TSE in combination with data on the costs of specific survey procedures.

Within a sample of respondents representing the population of interest, TSE recognizes that survey research seeks to measure accurately particular constructs or variables. For a specific survey, resulting measures might deviate from this goal due to four error categories:

- Sampling errors
- Nonresponse errors
- Coverage errors
- Measurement errors

The TSE framework explicitly considers each of these potential error sources and provides guidelines for making decisions about allocations of available resources. The result is that the sum of these four error sources (the total survey error) can be minimized for estimates developed from survey data.

The subsequent sections contain discussions of each error type and its relevance to EM&V for energy-efficiency programs. This chapter also describes current best practices for identifying, measuring, and mitigating these errors.

2.2 Sampling Errors

Sampling errors are random errors resulting from selecting a sample of elements from the population of interest, rather than from conducting a census of the entire population of interest. For practical or monetary reasons, it is often necessary to use a sample relative to an entire population. Although differences will likely occur between the sample and the population, so long as the sample has been based on probability sampling methods, these differences will likely be insubstantial.

A sampling error is the TSE component that is most frequently estimated, using measures such as the standard error of the estimate. Two methods commonly used to reduce sampling error are these: increasing the sample size; or ensuring the sample adequately represents the entire population. (Sample designs, sampling errors, confidence intervals and precision of estimates, and sample selection are discussed in a separate chapter.)

2.3 Nonresponse Errors

For any survey, some sampled customers likely will not complete the survey. Consequently, nonresponse error may occur if the nonrespondents differ from the respondents on one or more variables of interest. Nonresponse error may also occur when respondents fail to answer individual questions or items in the survey. Note that “nonresponse” is not necessarily the same as “nonresponse bias.” Such bias occurs when differences emerge between respondents and nonrespondents on one or more measures important to the analysis of gross energy savings.

For energy-efficiency EM&V surveys, the salience of the topic likely corresponds to the survey response rate (that is, interested individuals are more likely to respond). Consequently, nonresponse bias should be treated as a potential issue in designing survey implementation procedures.

2.3.1 Best Practices for Minimizing Nonresponse Errors

The following techniques have proven effective in reducing nonresponse among various target audiences:

- **Reduce the respondents' costs in completing surveys.** This is done by building trust and legitimacy in the respondents' eyes and by convincing the respondents they will receive a benefit from responding. The tools for this include advance letters, follow-up attempts, extending the data collection period, and incentives.
- **Highlight sponsorship of the survey** when it involves an organization with high credibility among the respondents, such as an electric or gas utility, a regulatory commission, a state or federal agency (Department of Energy), or a respected non-governmental organization. Having a credible sponsor usually increases the response rate.
- **When surveying organizations, identify appropriate respondents** to report on an organization's behalf. Then appeal to that individual to respond as the organization's representative. If a superior in the organization identifies an individual as the designated respondent, cite the superior when corresponding with the target respondent.
- **Avoid defining specific survey topics when introducing the survey** to sampled customers. Rather, describe the survey in terms as general as possible to reduce the likelihood of respondents making selections by their interest in a topic.

The potential for nonresponse bias can be estimated using these methods.

- **Collecting data (often a subset of survey questions) from nonrespondents** offers the most direct measure of nonresponse bias, although it can be difficult to obtain a representative sample of nonrespondents.
- **Comparing the responses of early responders (responders on the first contact) with those of responders who are more reluctant or difficult to reach.** This strategy assumes similarities between nonrespondents and reluctant or hard-to-reach respondents.

Where the potential for nonresponse bias has been identified, it is possible to weight the data to attempt to correct for underrepresentation of specific segments of the population. For example, where characteristics of the population are known, sample weights can be developed to adjust the proportion of these characteristics in the sample to match the characteristics of the population. Even when sample weights are used to adjust for nonresponse, however, the researcher has no assurance that the results account for differences between the individual respondents and nonrespondents from a particular segment.

2.4 Coverage Errors

When a sample (even a probability sample) excludes certain members of the population of interest, coverage errors may occur due to differences between the portions of the population excluded and the remainder of the population. A common example of this is a telephone survey that omits households without landlines. This also occurs in surveys of organizations that are selected based on their SIC codes, since new businesses may not have been classified yet and some businesses may have been classified incorrectly. Non-coverage might also result from the exclusion of some population members due to the following: geographic areas, language differences, physical challenges impairing the ability to respond, and individuals living in institutions.

An issue currently faced when using general population telephone surveys is the increasing number of households without landline telephones—recently estimated at more than 30% of all U.S. households (Blumberg, 2011). The likelihood of a household being "wireless only" relates to a number of demographic characteristics, such as these:

- Age (younger adults are less likely to have landlines);
- Household types (unrelated adults living together are more likely to be wireless);
- Own/rent status (renters are more likely wireless); and
- Household income (adults living in poverty are more likely wireless).

Further, the study indicated that one in six adults in the United States receives most or all telephone calls on wireless phones, even though there is a landline telephone at the residence. These data suggest telephone survey samples that do not include wireless phone numbers may produce data subject to "coverage error." (However, for surveys of program participants in which customers provided contact information, the chance of coverage bias due to missing cell phone-only households is reduced.)

A related issue is the "do not call" list maintained by some utilities. Customers who have requested that they not be contacted regarding certain matters are a potential source of coverage bias for energy-efficiency surveys.

2.4.1 Best Practices for Minimizing Coverage Errors

The following techniques have proven effective in reducing nonresponse among various target audiences:

- Evaluate the sample frame carefully to determine whether the listings match populations of interest. In your review, consider these questions: (1) Is the list up to date? (2) Are telephone numbers or other contact information current? (3) Does the list include wireless and landline phone numbers?
- Use dual sampling frames for general population surveys. For example, use cell phone number samples in addition to directory-based (land-line) samples.
- Define the population accurately for which the survey results are appropriately generalized. Thus, any segments not covered in the sample frame are clearly identified.

2.5 Measurement Errors

For most surveys, measurement error presents the most common and problematic error type. The term "measurement error" covers all biases and random variance arising when a survey does not measure its intended target. (This discussion does ***not*** include random errors, where respondents might answer a question differently over repeated trials. That results in increased variance, but not bias.)

In this chapter, measurement error is described as a systematic pattern or direction in differences between respondents' answers to a question and the correct answer. Such error occurs during data collection, rather than from sampling, nonresponse, coverage, or data processing. For example, respondents tend to over-report behaviors they believe are looked upon favorably and underreport behaviors they believe are viewed unfavorably (social desirability bias).

Measurement error results from the following factors:

- Respondent behaviors or responses to questions;
- Interviewers' influence on respondents; answers (interviewer effects);
- Question and questionnaire design; and
- Survey method of administration (mode).

The next sections describe how each of the first three measurement error sources can affect data quality and the best practices for reducing these effects. At the end of this section is a list of best practices for minimizing measurement errors. The effects of survey administration methods on measurement error are discussed in the chapter titled, "Survey Administration (Mode) Considerations."

2.5.1 Respondent Behaviors and Responses

Social desirability, acquiescence bias, and recall errors present the three most relevant bias sources, based on respondent behaviors.

2.5.1.1 Social Desirability Bias

This refers to the tendency of respondents to misreport their attitudes or behaviors intentionally in ways that make them seem appear to be doing "the right thing" in the eyes of interviewers or researchers. For example, in more than 50 years of behavior studies on voting, survey respondents have consistently reported voting at a higher rate than the turnout at the polls has actually indicated. Similarly, as energy-efficiency actions are widely viewed as socially desirable behaviors, it is expected that some respondents will over-report that they engaged in energy-efficient behaviors or that would have purchased an energy-efficient appliance, even had a rebate not been offered.

Voting behaviors provide a common focus for the study of socially desirable responding, as a well-established measure exists (official records of voter turnout) against which voting self-reports can be validated. However, no such validator exits for measures designed to determine whether a respondent would have purchased an energy-efficient appliance without an incentive. Thus, for questions about energy-efficiency actions and behaviors, wording that legitimizes

socially undesirable behavior can be used to mitigate social desirability bias. (This strategy has also been shown to reduce social desirability bias in surveys of voting behavior.)

For energy-efficiency surveys, a question measuring self-reports of energy-efficiency actions taken by respondents might be worded as:

> We often find that people have not done things to reduce energy use in their homes. They aren't sure how to do them, they don't have the right tools, or they just haven't had the time. For each of the following activities, please tell me if you have done this in your home. (Holbrook, 2010)

Social desirability bias primarily emerges as an issue for interviewer-administered surveys. Consequently, removing the interviewer's presence for self-administered survey modes reduces the pressure for socially desirable responding.

2.5.1.2 Acquiescence Bias

This refers to the tendency for respondents either to: (1) select an "agree" response more often than a "disagree" response; or (2) select a positively-worded response category more often than a negatively-worded response category, regardless of a question's substance.

In several studies using split-sample question wording experiments, Schuman and Presser (1996) demonstrated a classic example of acquiescence bias. They consistently found a difference between the percentage of respondents selecting the "agree" response when asked to agree or disagree with this: "Most men are better suited emotionally for politics than women." This wording received a higher "agree" rate than did the question, "Would you say that most men are better suited emotionally for politics than are most women?"

When respondents were presented with a forced choice question in other response categories indicating that men and women were equally suited or that women were better suited than men in this area, the result was a consistently lower agreement rate. For questions asked in the agree/disagree format, the percentage of responses indicating men were better suited for politics was consistently from 10 to15 percentage points higher than the results of the forced-choice format.

In questions asking about energy-efficiency actions, acquiescence bias is expected when statements are worded in a positive direction.

2.5.1.3 Recall Errors

These present another potential bias source based on respondent behaviors. Survey questions often ask respondents to recall specific events or to report on the frequency with which they have engaged in certain behaviors. Cognitive scientists and survey researchers have identified three factors correlating with errors in recall of retrospective events or behaviors:

- **Intervening events.** As related events occur—or new information related to the original event is obtained—individuals may lose the ability to recall accurately the specific details of any one event.
- **Recall becomes less accurate** with the passage of time.

- **Salient events are remembered more accurately** than less-salient events.[2] For energy-efficiency evaluations, the length of a recall period can be an important element in estimating gross energy savings. Respondents typically are asked to recall whether an event (such as purchase of an energy-efficient appliance) or the frequency of a behavior (such as the number of CFLs purchased) occurred within a specified time period.
- **Recollections of relatively infrequent events, such as purchases of a major appliance, are subject to telescoping errors.** That is, the events may have occurred earlier or later than was reported. Respondents purchasing a major appliance relevant to the survey but outside of the specified timeframe may report the event as occurring within the timeframe.
- **Recall decay—**the inability of respondents to recall events or frequencies of behaviors**—**tends to affect the accuracy of a respondents' recall of the frequency of relatively routine events (such as the number of CFLs purchased in a specific period).

2.5.2 Satisficing

One way respondents may introduce measurement error into their responses is by "satisficing"—taking actions enabling one to meet the minimum requirements for fulfilling a request or achieving a goal. When a survey question requires a great deal of cognitive work, researchers have found that some respondents use satisficing to reduce that burden. (Krosnick, 1991) The following behaviors have been observed in respondents attempting to reduce the amount of cognitive effort involved in responding to a survey:

- Choosing "no opinion" response options frequently when it is offered.
- Using the same rating for a battery of multiple objects rated on the same scale.
- Tending to agree with any assertion, regardless of its content (acquiescence bias).
- Choosing socially desirable responses.

Satisficing tends to occur in questions designed to measure knowledge, attitudes, and self-reports of behavior. The likelihood of respondents' engaging in satisficing is associated with respondents' cognitive abilities, motivations, and task difficulties.

2.5.3 Interviewer Errors and Effects

In interviewer-administered surveys, the interviewer's presence can negatively influence the quality of survey data in several ways, as noted below and in the extensive literature addressing interviewer errors and effects in sample surveys: (Biemer, 1991):

- As an interview is a social interaction, both the observable characteristics of interviewers and the manner in which interviewers interact with respondents can influence responses to survey questions.

[2] Eisenhower, 1991

- Interviewers can administer surveys differently to different respondents. For example, interviewers may do the following: (1) fail to follow skip patterns correctly, (2) ad lib or change the wording of specific questions, or (3) falsify data.
- In response to respondents' questions or difficulties, interviewers may probe or offer assistance in ways that affect respondents' answers.

The use of telephone interviews and self-administered surveys eliminates some potential effects related to social interactions between interviewers and respondents. Interviewer training—especially training that entails monitoring performance during interviews—provides the most effective way to identify and address potential sources of interviewer errors and effects.

2.5.3.1 Questionnaire and Question Design

Researchers tend to view questionnaires and questions as measurement devices, eliciting information from respondents. As a result, respondents' perspectives are frequently overlooked when questionnaires and questions also serve as a source of information for respondents to draw upon as they provide useful, informative answers to questions asked. (Schwartz, 1999)

Both the questionnaire (layout, formatting, and length) and the questions (wording, response categories, and context and order of questions) present information to respondents and, thus, can affect responses.

2.5.3.1.1 Questionnaire Length

It is commonly known that the longer the questionnaire, the more likely it is that respondent fatigue or loss of concentration becomes an issue. However, the answer to the question, "How long is too long?" differs for different survey modes and topics. The interviewer's skill is also a critical factor in terms of developing rapport with a respondent and maintaining the respondent's motivation.

In general, long surveys can be completed most successfully through personal interviews, while telephone surveys are most likely to be completed successfully when they are short. There is less of a consensus on the effect of questionnaire length for self-administered surveys (mail and Internet). Some research suggests that self-administered survey modes, especially Internet surveys, need to be relatively short to prevent respondents from abandoning the survey before it is completed. However, experience has shown that long self-administered surveys (ranging from 20 to 30 minutes) can be successfully administered, especially for mail questionnaires.

2.5.3.1.2 Open-Ended and Closed-Ended Questions

Although the great majority of energy-efficiency evaluation survey questions are closed-ended, there are advantages using an open-ended format for certain questions. For example, some researchers believe that open-ended questions about quantities—such as the numbers of times a respondent visited a specific Website—produce less bias than closed-ended questions. Specifically, this tends to apply to grouped, closed-ended response categories, such as: "at least one time per week" and "one to three times per month."

Response categories for closed-ended questions convey information about researchers' expectations. Also, many respondents tend to avoid extreme (high and low) scale points.

However, an open-ended question for which response categories are not provided avoids potential data-quality issues.

Similarly, for questions addressing the relative importance of issues facing the country, the closed-ended response categories offered to respondents indicate the issues that researchers think are most likely to be mentioned. This reduces the likelihood of respondents addressing issues not on the list. Despite this, closed-ended questions are used more often, as they are easier to code, process, and analyze. A general rule for using closed-ended questions is to ensure the response categories are comprehensive. (Krosnick, 2009)

2.5.3.1.3 Respondents' Interpretation of Questions

Since respondents must understand questions being asked, the researcher must determine whether the respondents' understanding of the questions matches the researcher's intent. Even for a seemingly straightforward question (such as, "What things do you typically do in your household every day to conserve energy?"), it is important to have some knowledge of the respondents' typical tasks.

Differences tend to occur in the literal understanding of the question. (Schwartz, 1999) For example, although respondents are likely to understand the literal meaning of a question, they must still determine the types of actions or activities of interest to the researcher. Consequently, in surveys about energy efficiency, respondents may ask themselves questions such as:

- "Should I report turning off lights when I leave the room, or is that too obvious?"
- "If I have an automatic set-back thermostat, is that considered an everyday activity?"

For questions open to multiple literal interpretations, researchers can guide respondents by using common examples of the types of information sought.

2.5.3.1.4 Question Order

The order of questions in a survey affects responses. When answering a specific question, respondents are likely influenced by cues and information from previous questions. For example, previous questions can present a priming effect—making certain issues more salient. Asking about the importance of energy efficiency before asking respondents about their energy-efficiency behaviors likely implies that those behaviors should be consistent with respondents' stated views on the importance of energy efficiency.

2.5.4 Best Practices for Minimizing Measurement Errors

- **Use pretesting to identify potential measurement errors**, such as instances in which respondents either misinterpret a question or are unable to provide an accurate answer.
- **Use salient events or dates in recall questions** to mark the relevant time period (bounded recall). Where possible, reduce burdens on respondents by shortening the recall periods.
- **Word the questions carefully** so respondents understand it is permissible to report engaging in non-socially desirable behaviors.

- **Use cognitive interviewing** as part of the survey pretest to explore how respondents interpret the questions and construct responses (Madans, et al. 2011).
- **To minimize acquiescence bias, avoid "agree/disagree" questions.** Instead, use questions explicitly presenting positive (agree) and negative (disagree) responses in the question stem, such as: "Would you say that most men are better suited emotionally for politics than are most women, that men and women are equally suited, or that women are better suited than men in this area?"
- **Use multiple-item measurement scales when assessing attitudes or reported behaviors**, and pre-test these scales to ensure unidimensionality and internal consistency. A multiple-item measurement scale consists of a number of individual questions combined into a single value. Using multiple-item measures usually increases the reliability of the measure.
- **Train interviewers and monitor the quality of their work** through observational interviews to reduce interviewer errors and interviewer effects.

2.5.5 Best Practices for Measuring Self-Reports of Behaviors

Evaluations of energy-efficiency programs often use self-reports of energy-efficient behaviors (or behavioral intentions). Thus, self-report surveys are designed to: (1) identify barriers in achieving gross energy savings; and (2) help explain differences in energy consumption between treatment and control group customers in programs with experimental designs. The best practices for these surveys of attitudes, behaviors, and behavioral intentions are described in the following sections.

2.5.5.1 Multiple Item Measurement Scales

Since the 1930s, survey researchers have used multiple-item scales to measure attitudes or reported behaviors. Based in psychometric theory, the rationale for multiple-item, self-reported behavior measurement suggests four primary advantages:

1. A set of multiple items can represent the construct (attitude or behavioral report) more completely than can a single item.
2. Combining items reduces potentially idiosyncratic influences of any single item.
3. Aggregating across items increases the reliability (or precision) of measures.
4. Using multiple items more finely distinguishes among respondents, potentially providing a measurement scale appropriately treated as continuous. (Nunnally, 1978)

In many cases, multiple-item scales of attitudes or self-reported behaviors treated as interval-level or continuous variables (item 4 in the list above) present important implications for statistical analyses of these data. Measures of central tendencies or dispersions prove appropriate for interval or continuous variables, and relative differences in scores between groups of respondents can be calculated. Multiple-item scales also produce variables well suited for use in regression models estimating gross energy savings.

Two procedures have allowed the development of summated multiple-item measures:

1. Factor analysis to verify multiple items measuring a single underlying construct (unidimensionality); and
2. A measure of internal consistency using Cronbach's alpha (coefficient of reliability) or a similar measure of the internal consistency of the measurement scale.

3 Developing Questions

To measure respondent self-reports of attitudes or behaviors in closed-ended questions, the design of the questions entails decisions about these critical elements:

- The order of response categories to be presented to respondents;
- The use of a rating or ranking scale;
- The type of rating scale; and
- The use of a middle or neutral category in a rating scale.

A summary of current evidence and best practices for each of these decisions is discussed below.

3.1 Order of Response Alternatives

The responses to closed-ended questions can be influenced by the order in which response categories are presented. For self-administered questionnaires and "show cards" used in personal interviews—where response categories are presented visually—research has shown a primacy effect often occurs. That is, respondents tend to select the answers offered early in the list. However, where response categories are presented verbally by an interviewer (whether on telephone or in person), a recency effect tends to occur, where respondents select answers offered later in the list. (Sudman, 1996) These research findings demonstrate the need to rotate the order of response alternatives offered to respondents.

3.2 Rating or Ranking?

Although rating scales commonly are used in energy-efficiency evaluation surveys, some situations have shown ranking to be a more effective method for measuring the importance of a specific issue or behavior. When the primary goal for a question is to determine the order of two or more objects, a ranking format may be most useful. (Visser, 2000)

3.2.1 Use of Ranking Scales

Ranking scales avoid the problems of non-differentiation, which occur when rating scales produce very similar ratings for a set of objects. However, rating scales are more commonly used in energy-efficiency evaluation surveys for the following reasons:

- Ranking is a more cognitively difficult task for respondents to complete, especially when dealing with a relatively large number of items; and
- Ranking scores prove more difficult to analyze. (As no assurance exists of equal distances between rankings, they cannot be used appropriately as interval measures.)

3.2.2 Use of Rating Scales

As previously mentioned, rating scales are the predominant method used for measuring self-reports of attitudes or behaviors. The basic types of these scales are classified as:

- Bipolar (from negative to positive, with a neutral point in the middle); or
- Unipolar (from a zero point to a highly positive point, such as a range from "no importance" to "extremely important").

After selecting the type of rating scale to use, the next decision is the length or the number of points on the scale. A quick review of questionnaires for energy-efficiency evaluations yields a wide range, from dichotomous (yes/no) scales to scales having as many as 100 points.

An important consideration in such decisions is whether to utilize scale points that divide the continuum into equal distances. If, for example, a scale offers a choice between "poor," "good," and "very good" but these choices have no numeric labels, then the continuum is not divided equally, as "good" and "very good" appear more closely related than "good" and "poor."

Scales using numerical labels meet the "equal interval" requirement. Many studies suggest data quality can be improved by labeling all scale points, rather than labeling only end points and neutral points. (Krosnick, 1999) Study findings indicate that applying these two techniques improves the results:

- Using words to anchor end-points and perhaps mid-points; and
- Using numbers to label each point on the scale.

As to the optimal number of scale points, reviews of research show the greatest measurement reliability results from seven-point scales for bipolar scales and five-point scales for unipolar scales.

3.2.3 Use of Middle Alternatives or Neutral Scale Points

Having a middle alternative (or a neutral alternative) increases the reliability of a measure, according to studies that examined the differences in reliability of an item's measurement —specifically, the use of a middle alternative in a scale. (O'Muircheartaigh, 1999) Some researchers advise using a middle category in a rating scale when a significant number of respondents are likely either to be uninformed or to have no opinion on the issue. Research also shows that the use of a middle alternative changes the frequency distribution of responses across all categories, but it often does not affect the ratio of responses on either side of the scales' middle point. (Schuman, 1981)

A recent alternative is to omit the middle category and then measure the intensity of the attitude. In this option, using a scale ranging from "strongly agree" to "strongly disagree" enables researchers to separate those who definitely hold a certain attitude from those who are simply inclined in a particular direction. (Converse, 1986) A number of experimental studies have shown data quality for a specific measure usually does not differ significantly, regardless of whether a neutral/no-opinion scale point is offered. (Schuman, 1996) In a 2002 study, Krosnick reported,

> The vast majority of neutral or no-opinion responses are not due to completely lacking an attitude, but are most likely to result from a decision not to do the cognitive work necessary to report it (satisficing), a decision not to reveal a potentially embarrassing attitude (social desirability bias), ambivalence, or question ambiguity.

This suggests the best practice for measuring attitudes or behavioral intentions entails omitting the neutral or no-opinion response category and encouraging respondents to report whatever opinion they have.

3.3 Summary of Best Practices for Question Design and Order in a Questionnaire

In their chapter on the design of questions and questionnaires, Krosnick and Presser advise the following when designing survey questions: (Krosnick, 2009)

- Use simple, familiar words, avoiding jargon, technical terms, and slang.
- Avoid words with ambiguous meanings; aim for words that all respondents interpret the same way.
- Use specific and concrete wording rather than general and abstract terms.
- Make response categories exhaustive and mutually exclusive.
- Avoid leading or loaded questions that push respondents toward an answer.
- Ask one thing at a time; avoid double-barreled questions.
- Avoid questions with single or double negations.

Further, Krosnick and Presser offer this advice regarding question order:

- To build rapport between respondents and researchers, make early questions easy and pleasant to answer.
- Questions at the beginning of a questionnaire should explicitly address the survey topic, as described to the respondent before the interview.
- Questions on the same topic should be grouped together.
- Questions on the same topic should proceed from the general to the specific.
- Questions on sensitive topics, which might make respondents uncomfortable, should be placed at the end of the questionnaire.
- Use filter questions to avoid asking respondents questions that do not apply to them.

3.4 Survey Administration (Mode) Considerations

The wide range of data collection modes available to survey researchers tend to fall into one of these categories:

- Interviewer-administered modes, such as personal or face-to-face interviews and telephone interviews; or

- Self-administered modes, such as mail or Internet surveys.

With advances in information and communication technologies, variations exist for each of the primary data collection modes. For example:

- Personal interviews can be conducted by an interviewer who records responses directly onto a laptop or electronic tablet.
- Self-administered questionnaires can be administered by audio-CASI, with questions recorded on an electronic device and played back to respondents, who enter responses electronically.
- Telephone interviews can be conducted by Webcam, in which respondents use either a voice-over Internet protocol or their phone keys to specify their answers.

The choices of data collection modes for energy-efficiency evaluations typically involve assessing strengths and weaknesses of a range of factors such as:

- Ability to access to a representative sample of the population of interest;
- Types of questions to be asked;
- Cost and time required for implementation; and
- Length, complexity, and content of the questionnaire.

3.4.1 Face-to-Face Personal Interviews

Considered by many survey researchers to be the "gold standard," face-to-face personal interviews generally result in high response rates, even for relatively long questionnaires (45 minutes or more). Through this approach, interviewers can manage complex questionnaires and those requiring visual or verbal background or explanations for the survey questions. However, face-to-face personal interview surveys are fielded less often due to their relatively high cost, as compared to other survey modes. Other key drawbacks are these:

- The longer time required to complete data collection;
- The logistical difficulty of quality control measures, such as observing interviewers' conducting the interviews; and
- The potential for interviewer effects resulting from the interviewer-respondent interactions.

3.4.2 Telephone Interviews

Telephone interviews have surpassed face-to-face personal interviews as the most common interviewer-administered survey mode for these reasons:

- The relatively lower cost per completed interview;
- The availability of off-the-shelf random-digit dialing (RDD) samples of the general population;
- The shorter length of time required to complete data collection; and

- The high proportion of households in the United States with a telephone.

With the advent of computer-assisted telephone interviewing (CATI), telephone surveys can accommodate complex questionnaires that apply skip patterns customized to respondent answers. Also, these interviews can be centrally monitored for quality control.

The key drawbacks of telephone interviews are these:

- The comparatively low (and declining) response rates;
- The relatively short time respondents can be expected to remain engaged (usually no more than 15 to 20 minutes);
- The increasing number of households using call-screening devices; and
- The increasing number of households without landline telephones.

Additionally, it is difficult to ask sensitive questions through telephone interviews, and social desirability bias presents a potential threat.

As a result of decreased coverage and response rates, telephone surveys are becoming less representative of the population of interest, except when mobile phone numbers are included in the survey. However, using listed samples of utility customers or program participants who have provided contact information can facilitate the contact of general-population households.

Note that when contacting a respondent by cell phone to conduct a survey, it is strongly recommended that the survey not be conducted if the respondent is driving a motor vehicle at the time of the call. In these cases, the interviewer should be instructed to make an appointment for a better time to call the respondent.

3.4.3 Mail Questionnaire Surveys

While the advantages of having an interviewer administer the questionnaire are noted above, there are also potential advantages for mail and self-administered questionnaires (without an interviewer). Self-administered questionnaires have been shown to: (1) produce more accurate or candid data for sensitive questions, and (2) reduce social desirability bias.

Mail questionnaires can be sent to anyone with an address. Also, respondents do not have to be home at any specific time, as is required for face-to-face personal interviews or telephone interviews. While completing a mail questionnaire survey, respondents can look up personal records, utility billing statements, or purchase information.

Although mail questionnaires often are described as the lowest-cost alternative among survey modes, this approach—in our experience—requires at least two follow-up mailings and, in some cases, relies on an incentive to increase the response rate. This increases cost of fielding the survey. Other drawbacks typically associated with mail questionnaire surveys are these:

- Relatively low response rates (in many cases, rate comparable to a telephone survey);
- Longer data collection periods;

- Skip patterns must be relatively simple to avoid confusing respondents;
- Loss of control over who answers the questions; and
- Loss of control regarding the order in which questions are viewed and answered.

3.4.4 Internet Surveys

Internet surveys have increased in popularity, especially as the percentage of households and individuals with access to the Internet has increased. These surveys offer the advantage of lower cost (no expenses for paper, printing, mailing, telephones, or interviewers). Further, once the fixed costs of programming and set-up have been incurred, a much larger sample size can be used—even internationally—with very small marginal cost increases.

Internet surveys usually require very short data collection times, with most responses received within one week, although follow-up contacts should be made with nonrespondents to increase response rates. Note, however, that coverage bias for potential respondents who do not have access to the Internet remains an issue with online surveys.

Consistency in the appearance of the survey is also an issue. While enhanced Internet survey software allows for complex skip patterns and sophisticated graphic, different hardware and software used by respondents can result in differences in a questionnaire's appearance and presentation.

As with mail questionnaire surveys, the absence of an interviewer requires that the questions be relatively simple and straightforward. Still, with Internet surveys, the respondents' willingness to answer sensitive questions candidly is increased, and the likelihood of social desirability bias is decreased.

3.5 Using Multiple Survey Modes: Mixed-Mode Surveys

In this century, a major trend in survey research has been the increased use of combined survey implementation modes. (Dillman, 2009) It has long been a practice to mix modes in:

- The survey's contact phase (for or example, using an advance letter to contact respondents for telephone surveys or face-to-face interviews); and
- Completing different portions of a survey.

What has been relatively new in survey research, however, is use of mixed-mode surveys in which some respondents provide data using one mode, while others provide data using a second (or third) mode. (Couper, 2011)

This section describes this relatively new approach to mixed-mode surveys. Their increasing use has been driven by several factors, including declining response rates, coverage problems in single-mode surveys, and the development of new survey modes—such as interactive voice response (IVR) and Internet-based methods.

Research has shown that mixed-mode surveys can achieve higher response rates and better coverage of populations of interest. As different methods have different strengths and weaknesses, using a variety of methods can provide complementary results. (de Leeuw, 2005)

Still, mixed-mode surveys present drawbacks—such as increased measurement error—because different survey modes can produce different responses to the same question. (Christian, 2008)

In a 2011 publication addressing questions about using a mixed-mode survey, Mick Couper cited two strategies in dealing with potential mode differences:

- The **unimode construction** approach constructs questionnaires to be as identical as possible.
- The **correction approach** entails accepting fundamental differences in data collection by different modes *and* designing the data collection instrument to maximize the benefits of each mode; statistical adjustments then are made across the modes used. (Couper, 2011.)

A third strategy is to combine these approaches when designing and implementing mixed-mode, energy-efficiency evaluation surveys. For example, in mixed-mode surveys using telephone and Internet, the fixed-page telephone interview survey—where respondents are asked questions in a specified sequence by CATI—can best be replicated by an Internet survey, where respondents see one question at a time, and cannot progress to the next question until the first is answered. Also, an IVR Internet survey can also be used to replicate the presence of an interviewer for such mixed-mode surveys.

For a mixed-mode survey using mail and Internet questionnaires, the scrolling-page Internet survey design best replicates mail questionnaire design, where respondents can turn ahead pages if they wish to see questions in the survey.

Replicating in two survey modes how questions are presented provides an opportunity to increase the effectiveness of energy-efficiency evaluation surveys, while increasing coverage and response rates. New technologies and advancements in survey research capabilities will continue to provide additional ways of mixing modes and to increase survey effectiveness and quality.

4 Minimum Reporting Requirements for Energy-Efficiency Evaluation Surveys

Survey research organizations—such as the American Association for Public Opinion Research (AAPOR) and the Council of American Survey Research Organizations—require their members follow appropriate professional guidelines for disclosing and reporting survey methods and findings. The goal of these organizations is to advance the state of knowledge and practice by providing sufficient information to permit review and replication by other researchers.

The AAPOR offers various guidelines regarding the minimum essential information on survey methods to be disclosed in research reports:

- Survey sponsor and the firm conducting the survey;
- Survey purpose and specific objectives;
- Questionnaire and exact/full wording of questions as well as any other instructions or visual exhibits provided to respondents;

- Definitions of populations under study;
- Descriptions of the sampling frame used to identify populations under study;
- Sample design, including: clustering, eligibility criteria and screening procedures, selection of sample elements, mode of data collection, and the number of follow-up attempts;
- Sample selection procedures (how sample cases were selected);
- Documentation of response or completion rates, numbers of refusals, and other dispositions;
- Discussion of the findings' precision, including sampling error, where appropriate;
- Descriptions of special scoring, editing, data adjustment, or indexing procedures used;
- Methods, locations, and dates of fieldwork or data collection; and
- Copies of interviewer instructions for administering the questions.

Following the disclosure and reporting guidelines available on the AAPOR Website serves to advance knowledge and the state of practice for energy-efficiency evaluation research and, ultimately, this results in better-quality data and better decisions on energy-efficiency programs.

5 Conclusion

This chapter has provided an overview of the current state of survey research regarding the evaluation of energy-efficiency programs through: (1) developing estimates of gross energy savings, (2) determining well market effects, and (3) identifying process issues. For each topic covered—summarized below—readers are encouraged to consult articles and books cited in the references section as well as other sources covering these topics in much greater depth:

- Sources of survey error, such as nonresponse, coverage, and measurement;
- Best practices for measuring self-reports of attitudes and behaviors;
- Best practices for question wording and question order;
- Selection of survey modes and use of mixed-mode approaches; and
- Minimum guidelines for reporting and disclosure of survey research.

6 References and Resources

American Association for Public Opinion Research. Website (www.aapor.org/Best_Practices1.htm).

American Statistical Association. 1980. *What Is a Survey?* Washington, DC.

Biemer, Paul, and Robert M.Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman (eds.) 1991. *Measurement Errors in Surveys*. John Wiley & Sons.

Blumberg, Dr. Stephen J., and Julian V. Luke. 2001. "Wireless Substitution: Early Release of Estimates From the National Health Interview Survey." Division of Health Interview Statistics, National Center for Health Statistics, Centers for Disease Control.

Bradburn, Norman, Seymour Sudman, and Brian Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires.* John Wiley & Sons.

Christian, L. M., and D. A. Dillman, J. D. Smyth. 2008. "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys." J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. Lavrakas, M. Link, and R. Sangster (eds.) Advances in Telephone Survey Methodology. Wiley-Interscience.

Converse, Jean M., and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire.* Sage Publications.

Couper, Mick P. 2011. "The Future of Modes of Data Collection." *Public Opinion Quarterly.* Vol. 75, No. 5, Pages 889-908.

de Leeuw, Edith D. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." Journal of Official Statistics, Vol. 21, No. 2, pages 233-255.

Dillman, Don A., and Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell, Jodi Berck, Benjamin Messer. 2009. "Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR), and the Internet." Social Science Research, Vol. 38, pages 1-18.

Eisenhower, Donna, and Nancy A. Mathiowetz, David Moganstein. 1991. "Recall Error: Sources and Bias Reduction Techniques." Biemer, Groves, Lyberg, Mathiowetz, and Sudman (eds.) *Measurement Errors in Surveys*. John Wiley & Sons.

Groves, Robert M., and Lars Lyberg. 2010. "Total Survey Error: Past, Present, and Future." Public Opinion Quarterly, Vol. 74, No. 5, pages 849-879.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. Wiley Series in Survey Methodology. Wiley-Interscience: New York.

Holbrook, A. L., and J. A. Krosnick. 2010. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item-Count Technique." Public Opinion Quarterly, Vol. 74, No. 2, pp. 328–343.

Krosnick, Jon A. 1991. "Response Strategies for Coping with Cognitive Demands of Attitude Measurement in Surveys." *Applied Cognitive Psychology*, Vol. 5, pages 213-236.

Krosnick, Jon A., and Stanley Presser. 2009. "Question and Questionnaire Design." James D. Wright and Peter V. Marsden (eds.) *Handbook of Survey Research (2nd Edition).* Elsevier: San Diego.

Krosnick, Jon A., and John P. Robinson, Philip R. Shaver, and Lawrence Wrightsman (eds.). 1999. "Maximizing Questionnaire Quality." *Measures of Political Attitudes*. Academic Press, San Diego.

Madans, Jennifer, Kristen Miller, Aaron Maitland, and Gordon Willis. 2011. *Question Evaluation Methods: Contributing to the Science of Data Quality.* John Wiley and Sons.

Nunnally, Jum C. 1978. *Psychometric Theory* (2nd ed.) McGraw-Hill, New York.

O'Muircheartaigh, Colm, and Jon Krosnick, A. Helic. 1999. "Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data." Paper presented at the American Association for Public Opinion Research Annual Meeting, St. Petersburg, Florida.

Schuman, Howard, and Stanley Presser. 1996. *Questions and Answers in Attitude Surveys*: *Experiments on Question Form, Wording, and Context*. Sage Publications.

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context.* Academic Press.

Schwartz, Norbert. 1999. "Self-Reports: How the Questions Shape the Answers." American Psychologist, Vol. 54, No. 2, pages 93-105.

Sudman, Seymour, and Norman Bradburn, Norbert Schwartz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco.

Visser, Penny S., and Jon A. Krosnick, Paul J. Lavrakas. 2000. "Survey Research." Harry T. Reis and Charles M. Judd (eds.) *Handbook of Research Methods in Social and Personality Psychology.* Cambridge University Press.

Wickman, A., and B. Warneryd. "Measurement Errors in Survey Questions." Social Indicators Research, Vol. 22, pages 199-212.