Think before You Do: The Importance of Survey Design in Program Evaluation

Tami Buhr, Opinion Dynamics Corporation, Waltham MA

ABSTRACT

Survey research is a critical piece of the evaluator's toolkit. Nearly all evaluations require survey data in some form. Process evaluations rely on surveys of program participants to understand the effectiveness of program design and operations. Many impact evaluations estimate net energy savings from participant self-reports about their decision to save energy. Baseline studies use survey data to estimate the current state of the market for appliances and equipment, while potential studies use surveys to estimate the potential for more energy efficient equipment.

Despite the prominent role that surveys play in evaluation, survey design is not often given the same level of attention as other needed skills. Although survey design is a field unto its own, it is commonplace for people with no survey training or experience to write evaluation surveys. An evaluation is only as good as the data on which it is based, whether that data is the result of an engineering analysis or a survey. A poorly constructed survey instrument can lead to many undesirable results.

In this paper, we provide a review of survey design literature and best practices. We also discuss the importance of survey testing and monitoring to help determine if, in fact, the respondents understand the questions asked—just because a respondent answers a question does not mean they comprehend the question. The information presented will be useful for utility managers who must review many survey instruments but often do not have the background to understand effective survey design.

Introduction

Survey research is a critical piece of the evaluator's toolkit. Nearly all evaluations require survey data in some form. Process evaluations rely on surveys of program participants to understand the effectiveness of program design and operations. Many impact evaluations estimate net energy savings from participant self-reports about their decision to save energy. Baseline studies use survey data to estimate the current state of the market for appliances and equipment, while potential studies use surveys to estimate the potential for more energy efficient equipment.

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. Although survey design is a field unto its own, it is commonplace for people with no survey training or experience to write evaluation surveys. An evaluation is only as good as the data on which it is based, whether that data is the result of an engineering analysis or a survey. A poorly constructed survey instrument can lead to many undesirable results.

A survey is a structured conversation. Like any conversation, word choice can have a large impact on understanding. People can and do interpret the same word differently. It is relatively easy to think of examples of miscommunication in everyday conversations because two people interpreted the same words differently. Conversational miscommunications are so common that they are often used as comedic devices such as the Abbot and Costello's classic "Who's on First routine.

Miscommunications in conversation can usually be cleared up with additional conversation. Miscommunications in surveys are more problematic. Interviewers are typically required to read survey questions exactly as written and are given specific instructions about what they can and cannot say when respondents do not understand a question.

Numerous studies have also shown that that a slight change in survey wording, can have a large impact on survey responses. One well-known change in wording that causes a large change in survey results is swapping the word "forbid" for "allow" in questions asking about public policies. This effect was found

sixty years ago and has been replicated numerous times. In 1940, 54% of Americans said they thought public speeches against democracy should be *forbidden* (see Table 1). When an independent sample of respondents was asked if such speeches should be *allowed*, a much larger number, 75%, said they should not be allowed, which is equivalent to forbidding these speeches. These same questions were asked again in 1976 with similar differences between the question formats though both versions showed the American public had become more supportive of free speech during the intervening 36 years.

Table 1. Impact of Slight Changes in Question Wording on Survey Results

Do you think the United States should *forbid* public speeches against democracy?			**Do you think the United States should *allow* public speeches against democracy?**		
	1940	**1976**		**1940**	**1976**
Yes (forbid)	54%	21%	No (not allow)	75%	48%
No (not forbid)	46%	79%	Yes (allow)	25%	52%

The forbid/allow example is just one of many that illustrate the importance of careful survey design and testing when fielding a survey. In this paper, we provide a set of research based best practices for survey design that evaluators and utility managers can refer to when designing and reviewing surveys. We also discuss the importance of monitoring surveys to help determine if, in fact, the respondents understand the questions asked—just because a respondent provides an answer to a question does not mean they comprehend the question.

An expert in the field of survey design said that 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," (Fowler 1995, 2). Achieving this common level of understanding is not as easy as it seems. We provide guidelines to help evaluators and utilities know that their survey questions are actually measuring the concepts they are intended to measure.

General Principles of Question Wording

Most survey respondents are somewhat reluctant to participate in the survey and need a bit of convincing. Few energy evaluation surveys provide incentives to respondents to encourage them to participate so the convincing is done by the interviewers' polite requests. Even if respondents do receive an incentive, the payment is generally a token of appreciation and not an attempt to truly compensate respondents for their time. The only reward most receive is the psychological benefit people get from helping others.

As a result, survey respondents may not be highly motivated to listen to each question carefully and provide the most accurate response. Research has shown that many survey respondents take short cuts when answering questions by basing their responses on the most easily accessible information in memory (Krosnick 1991; Simon 1957). This is a process known as "satisficing". Respondents who find the questions difficult to understand or the respondent lacks the necessary information or knowledge to answer the questions are more likely to satisfice when answering survey questions. Respondents with less education or less investment in a subject are more likely to satisfice when answering survey questions.

The general question design principals outlined below will aid evaluators attempting to minimize satisficing.

1. **Ask questions that are clear and specific in what they are asking.**

2. **Ask about one subject at a time.**
Questions that ask about more than one subject are known as "doubled-barreled". Respondents may have different opinions about the different subjects in the question but are only allowed to give one. Evaluators cannot know which part of the question the respondent has answered when analyzing the question results.

> Double-Barreled Question:
> *Are you satisfied or dissatisfied with the audit program sign-up process and the time it took to schedule your audit?*
>
> Separated into Two Questions:
> *Are you satisfied or dissatisfied with the audit program sign-up process?*
> *Are you satisfied or dissatisfied with the audit program the time it took to schedule your audit?*

3. **Avoid the use of double negatives in questions.**
Questions with double negatives can confuse respondents. An example from a 2006 Gallup Poll shows the impact of question wording with a double negative:

> *Would you favor or oppose a bill that would **prevent** any foreign-owned company from owning cargo operations at seaports in the United States?*
>
> *Favor: 38%*
> *Oppose: 58%*
> *No Opinion: 4%*

The same question is reworded from negative to positive with a different result:

> *Would you favor or oppose a bill that would **allow** only U.S. companies to own cargo operations at seaports in the United States?*
>
> *Favor: 68%*
> *Oppose: 25%*
> *No Opinion: 7%*
>
> *Gallup Poll, March 13-16, 2006.*

4. **Response options should be exhaustive and mutually exclusive.**
Make sure that quantity ranges do not overlap such as the following:

> *How many CFLs did you purchase?*
> *0-5*
> *5-10*
> *10-15*
> *15+*

Questions should not require the respondent to pick the most applicable response among many. These are difficult questions for the respondent and unreliable. More than one response option could apply to a respondent in the following question:

> *Which of the following statements best applies to you? At the time I first heard about the rebate for the central air conditioner…*
>
> *I was already thinking about purchasing a central air conditioner*
> *I had already been collecting information about a central air conditioner*
> *I had already selected the central air conditioner I was going to get*
> *I had already installed the central air conditioner*
> *I had not thought about purchasing a central air conditioner at all*

Open-Ended Versus Close-Ended Questions

Surveys typically contain a combination of open-ended questions and closed-ended questions, also known as forced choice. Open-ended questions allow respondents to answer the question in their own words while close-ended questions require respondents to select their response from a provided list. Each has pros and cons.

By allowing respondents to give an answer in their own words rather than fit their response into a set of predefined categories, evaluators can gain unexpected information. Open-ended questions also provide more rich and detailed responses than close-ended question.

However, open-ended questions are more difficult to administer and analyze. Answering an open-ended question requires more time and thought on the part of the respondent, which some respondents are not willing to do. Interviewers need to be trained to encourage reluctant respondents to provide answers. Interviewers also need to be able to accurately record the response.

Though open-ended responses can provide good quotes, typically the raw responses need to be coded into a limited number of categories for analysis. A coding scheme needs to be developed and coders trained and evaluated to assure responses are being accurately grouped into the defined categories. Coding takes time and adds to the survey cost.

A short-cut that people sometimes take is to ask an open-ended question and allow the respondent to provide their answer in their own words. The interviewer "field-codes" the response into pre-defined categories that are not read to the respondent. This approach can reduce analysis time and survey costs, but it is not recommended in most cases. The interviewer becomes the coder and considerable training is typically required for each question to ensure that all interviewers are coding the open-ended responses correctly and consistently. A check would be to have the interviewer type the verbatim response as well as fit it into the precoded categories, but this would lengthen the survey and could cause an inpatient respondent to terminate the interview.

Given the challenges and costs of asking open-ended questions, most survey questions tend to be close-ended. Rating scales are examples of common close-ended questions. Many previously used and tested close-ended questions exist that evaluators that can use as starting points for their surveys. It is more difficult to develop a list of response categories from scratch for a question that could be asked as an open-ended question. The evaluator needs to make sure that the response list is complete, exhaustive, but not too long to be read. Survey designers sometimes take a short-cut by reading just a few possible response categories but also allowing the respondent to provide an alternative response as an "other" if the read responses are not applicable. In this case, it is better to ask a true open-ended question. Respondents will tend to pick a

response from the provided list even if it isn't the most appropriate response and not bother to provide an alternative response.

Researchers can often get very different answers to the same question when it is asked as a close-ended question compared to an open-ended. A typical question that is often asked as an open-ended question in evaluation surveys asks program participants how they learned about the program. The program is marketed through a number of channels and these results could be used to see which were most effective. The open-ended question could read: "Where did you hear about the program?" with the interviewer recording the verbatim answers. This approach requires the respondent to remember every place they heard about the program unprompted and also to take the time required to remember. If the evaluator is interested in evaluating the specific channels through which the program was marketed, he should instead, ask about each of the channels. The table below compares the different question designs.

Table 2. Open-Ended Question Compared to Close-Ended

Open-Ended	Close-Ended
Where did you hear about the program? [Interviewer: record verbatim answer]	You might have heard about the program from a number of different information sources. Did you hear about the program or see it advertised on…
	…the utility web site?
	…through an insert that came with your utility bill?
	…from your contractor?
	…from family or friends?
	…the radio?

The close-ended question involves more questions but will not take much more time as open-ended questions take longer to administer. The close-ended question will prompt the respondent's memory and will likely come up with more accurate results.

Some subjects are still better asked as open-ended questions. Questions asking for a quantity are best left as open-ended rather than providing response ranges in which the respondent must fit a quantity. Research suggests that these ranges can introduce bias into the responses (Schwartz et al. 1985). If necessary, the evaluator can easily recode the responses later into categories for reporting.

Factual Questions and Recall Error

Many energy evaluation surveys ask factual questions about past behaviors. For example, lighting evaluations often include a survey about CFL usage with questions about purchase behavior. A typical question might be: "How many CFLs did you purchase in the past year". For respondents to accurately answer a question such as this, they need to have encoded the information in memory (Krosnick 2010). This seems like an obvious point, however surveys sometimes ask about subjects that respondents never bothered to encode in memory. Behaviors that are not salient are less likely to be encoded, and even when they are encoded, they are more likely to be recalled in error. The lighting survey that asks about CFL purchases assumes that respondents made a point of encoding facts about the bulbs purchased at the time of purchase. The purchase of CFLs is likely not that important to many people and could be subject to recall error.

Unfortunately, respondents will often do their best to answer these types of questions to appear knowledgeable or please the interviewer. It is difficult, if not impossible, for evaluators to identify recall

errors once the data has been collected. However, the techniques and guidelines listed below can help reduce recall error:

- Ask a longer question; note this does not mean a more complicated question. Introduce the question by letting the respondent know the subject of the upcoming questions. A longer question gives the respondent some additional time to think about their answer.
- Encourage respondents to take their time and stress the importance of collecting accurate information.
- Select a reference time frame that is appropriate for the saliency of the question subject.
    - Shorter time frames should be used for less salient subjects. For example, a year is likely too long when asking about CFL purchases.
    - Evaluators should beware of selecting too short of a time frame as that can encourage respondents to report behaviors that happened earlier. A month is likely too short for CFL purchases as people might recall purchases made relatively recently, such as two months ago, but recall making the purchase even more recently when asked about the past month.
- Use aided-recall techniques to prompt the respondent's memory.
    - Provide cues to help respondents remember the situation referenced in the question. This may involve asking additional questions. The lighting evaluator may only be interested in CFL purchases, but the respondent may give a more accurate answer if asked to think about all lighting purchases. Table 3 shows the difference in strategy.

Table 3. Unaided Versus Aided Recall Question Series on CFL Purchases

Unaided Recall	Aided Recall
How many CFLs have you purchased in the past three months?	Next, I'd like to ask you some questions about any lighting purchases you may have made in the past three months. Please take your time in answering these questions. It is very important that I record the most accurate information possible.
	Have you purchased any light bulbs in the past three months?
	[If purchased bulbs ask:] Did you purchase any incandescent light bulbs in the past three months?
	[If purchased bulbs ask:] Did you purchase any compact fluorescent light bulbs in the past three months?
	[If purchased incandescent bulbs ask:] How many incandescent light bulbs did you purchase in the past three months?
	[If purchased CFLs ask:] How many CFLs did you purchase in the past three months?

Note: The different types of lighting should be described to the respondent at some point in the survey. These questions assume that earlier questions contained these descriptions.

Question Scales

Although rating scales are one of the most common question forms, there are so many different types and lengths of scales in use that it is difficult to know the best scale to use in a given situation. Fortunately, a large body of research exists on optimal scale use and design that can help evaluators select the best scale for their survey needs.

Bipolar versus Unipolar Scales

The first decision when writing a rating scale question is whether to use a bipolar or a unipolar rating scale. A bipolar scale measures both the direction and intensity of the attitude with the end points representing equally intense but opposite ends of the spectrum. A unipolar scale measures just intensity. A bipolar concept can be measured using either a bipolar or unipolar scale though one may be preferable to another. This is not true of unipolar concepts that can only be measured with a unipolar scale.

Some concepts are bipolar. A bipolar concept has both negative and positive sides. Often, there is a mid-point or neutral position that can also be measured (more on offering a neutral position below). A satisfaction scale is an example of a bipolar concept. A program participant may be dissatisfied or satisfied with his program experience. He may also have varying degrees of dissatisfaction or satisfaction. Or he may feel neutral about his experience with the program (see Figure 1).

Figure 1. Bipolar Concept Measured on a Bipolar Scale

1	2	3	4	5
Very Dissatisfied	Somewhat Dissatisfied	Neither Satisfied nor Dissatisfied	Somewhat Satisfied	Very Satisfied

Satisfaction can also be measured using a unipolar scale that measures degree of satisfaction ranging from “not at all satisfied” to “extremely satisfied” (see Figure 2).

Figure 2. Bipolar Concept Measured on a Unipolar Scale

1	2	3	4	5
Not at all Satisfied	Not Very Satisfied	Somewhat Satisfied	Very Satisfied	Extremely Satisfied

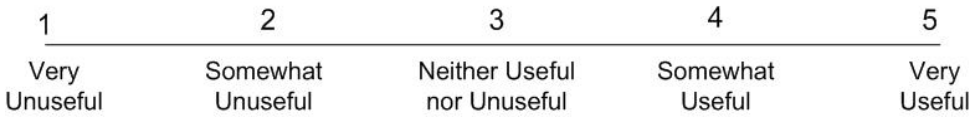
A unipolar concept has just one dimension and varying degrees of intensity. The concept is measured on either the negative or positive side of a bipolar scale, but not both. Usefulness is a unipolar concept. Participants in a program that provides customers with information about how to save energy may find that information useful or not useful with varying degrees of utility in between (see Figure 3).

Figure 3. Unipolar Concept Measured on a Unipolar Scale

1	2	3	4	5
Not at all Useful	Not Very Useful	Somewhat Useful	Very Useful	Extremely Useful

Participants cannot, however, find the information unuseful (see Figure 4).

Figure 4. Unipolar Concept Measured on a Bipolar Scale



Knowledge is another unipolar concept. The information program could increase participants' knowledge, but it cannot have a negative impact on knowledge. Program impact could range from not at all knowledgeable to very knowledgeable, but knowledge would not range from very unknowledgeable to very knowledgeable.

Before writing the scales, evaluators should determine whether their concept is one that can be measured on a bipolar or unipolar scale. Does the concept have both a negative and a positive side? Even if the concept can be measured using a bipolar scale, a unipolar scale may be the better choice. Research shows

Optimal Number of Scale Points

Another question facing survey designers is how many points to include on a rating scale. For a scale to provide a reliable and valid measure of a concept, respondents must uniformly understand the meaning of rating scale response categories. Scales with a small number of points are easier for respondents to understand so that respondents tend to interpret the categories in the same manner. The drawback of these scales is that they do not allow finer distinctions in attitudes and behaviors that most respondents are able to make. But scales with too many categories can only provide this higher level of distinction if each point has a clear and distinct meaning. Long scales without clear meaning can create measurement error.

The optimal number of scale points to maximize reliability and validity of survey responses has been the subject of numerous studies. The general consensus is that scales with a moderate number of points – five or seven – tend to have greater reliability and validity than scales with fewer or more points. Five-point scales are best for unipolar concepts while seven-point scales are best for bipolar concepts (Green & Rao 1970; Lissitz & Green 1975; Malhotra, Krosnick, & Thomas 2009).

Studies also show that associating descriptive labels with all numeric rating values and not just descriptors for the end-points increases the reliability of survey responses (Krosnick & Berent 1993). Internet and mail surveys can easily provide labels for all categories. For phone surveys, it is possible to provide verbal labels for scales of five or less with little problem, but it is more difficult with seven-point scales and is impractical for scales with more than seven points. Evaluators could provide labels on all scale points by breaking a seven-point scale into two questions with the first question asking direction (e.g. Overall, are you satisfied, dissatisfied, or neither satisfied nor dissatisfied with your experience with the program?) and then asking a follow-up that measures intensity (e.g. very satisfied, satisfied, slightly satisfied). Research suggests that this type of branching question does not take more time to administer and provides more reliable results compared to unbranched and partially labeled questions (Krosnick & Berent 1993).

Table 4. Branched Versus Unbranched Question

Single Unbranched Question
Using a seven point scale that ranges from 1 to 7 where 1 represents very dissatisfied and 7 represents very

satisfied, how satisfied are you with your overall program experience? *Very Dissatisfied 1 2 3 4 5 6 7 Very Satisfied*
Branched Question Series
Thinking about your overall experiences with the program, are you satisfied, dissatisfied, or neither satisfied nor dissatisfied? *1 Satisfied 2 Dissatisfied 3 Neither Satisfied nor Dissatisfied*
If Satisfied…
Are you very satisfied, somewhat satisfied, or slightly satisfied? *1 Very Dissatisfied 2 Somewhat Dissatisfied 3 Slightly Dissatisfied*
If Dissatisfied…
Are you very satisfied, somewhat satisfied, or slightly satisfied? *1 Slightly Satisfied 2 Somewhat Satisfied 3 Very Satisfied*
Questions are combined to produce an overall result on a seven-point scale: *Very Dissatisfied 1 2 3 4 5 6 7 Very Satisfied*

Providing a Middle or Neutral Position on Question Scales

Evaluators face another decision when constructing rating scales: whether to offer a "middle" or neutral position. Studies show that more respondents will choose a middle position when it is offered than if they have to volunteer that response. Research is mixed on whether not offering a middle alternative impacts the conclusions one would draw. Some studies show that respondents who would select a middle position end up selecting the two alternative sides of a scale in equal proportion (Schuman & Presser 1981). Other research shows that respondents who select a middle alternative would not necessarily answer the question in the same way as the other respondents in the survey so that the survey results are impacted. That is, removing a middle alternative shifts more people to one side of an issue than the other impacting the overall results (Bishop 1987). Given the mixed results, survey designers should be guided by the subject matter of the question. If a middle or neutral position is legitimate response on a question, survey designers should offer the response option.

Don't Know Responses

In most phone surveys, it is standard practice to not read a response of "don't know" to respondents. If the respondents say they don't know the answer, interviewers are typically trained to encourage respondents to provide an answer before recording the response as "don't know". This practice is based on the belief that people satisfice when answering survey questions so that responding "don't know" is easier than expressing an opinion. Research supports this idea. Studies show that respondents who are encouraged to provide an answer after initially saying "don't know" gave an opinion that correlated with their answers to other questions in a predictable manner (Krosnick et al. 2002). Likewise, respondents who said they did not know when asked a question testing their knowledge of a subject, were more likely to give a correct than incorrect answer when pressed for a response (Mondack & Davis 2001).

Not offering a "don't know" response only works if survey designers use branching and question skips to ensure that respondents are only asked questions that are appropriate. For example, a survey evaluating a residential HVAC program might contain questions about the contractor and what role he played in helping promote the selection of energy efficient equipment. Contractors sometimes provide materials that customers can read on different equipment options. It would not make sense to ask a program participant how much they learned from this type of material without first finding out if they even received

this material. If they did not, the survey should skip those respondents over any subsequent questions about the material. If respondents who were never given additional materials are skipped, it is appropriate to encourage other respondents to provide an opinion on the materials they did receive.

Social Desirability Bias

People naturally like to appear to be good, upstanding members of the community and will sometimes distort their survey response to give that appearance. Survey questions about behaviors or attitudes that are socially desirable or undesirable can suffer from response bias. For example, surveys that askpeople whether they voted in a recent presidential election typically find that 15% more claim to have voted than actually did compared to official turnout and verification of voting records.

Energy evaluation surveys that ask about energy efficient behaviors and attitudes could also be subject to social desirability bias. Not all respondents will feel that these behaviors are socially desirable, but an increasing number may as energy efficiency becomes more widely accepted. It is ironic that utility programs that successfully change attitudes and behaviors about energy efficiency may also make it more difficult to accurately measure program impact as customers will feel reluctant to report that they did not attempt to take actions.

Evaluators who want to minimize social desirability bias should consider wording their questions to make it more socially acceptable to not take energy efficient actions. Pollsters attempting to reduce over-reporting of voting will often preface the actual question with an introduction such as, "In talking with people about elections, we often find that many people were not able to vote because they weren't registered, were sick or just didn't have time. How about you? Were you able to vote in the presidential election this past Tuesday?" The goal is reassure respondents that it is okay if they have not voted so they feel more comfortable admitting it. Table 5 provides an example of a traditionally worded question about energy efficient behaviors and one that attempts to reduce over-reporting of these behaviors.

Table 5. Branched Versus Unbranched Question

Standard Question Wording	Wording to Reduce Social Desirability Bias
I'm 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 have done. How about…	In talking with people about ways they might reduce the amount of energy their household uses, we often find that people aren't sure what actions will save them energy or they are too busy make changes in their routines. I'm going to read you a list of actions you might have taken to save energy. After I read each one, please tell me if this is something you have been able to do in your household.
Turning off the lights when no one is in the room? Washing your clothes in cold water?	

Evaluators and program implementers may be understandably reluctant to reassure customers that not taking energy saving action is okay. It is important separate the evaluation effort from program marketing campaigns where such reassurances would be inconsistent with the purpose of the program. The evaluation touches a very small number of utility customers and will not impact program results. However, a biased evaluation of program effectiveness could impact future program success.

Self-administered surveys are also less likely to suffer from social desirability bias than telephone or in-person surveys. Internet surveys generally show less bias than telephone surveys on experiments. Internet

surveys are a good choice for surveys on subjects that may have social desirability bias if email addresses are available.

Survey Development and Testing Techniques

Before survey fielding begins, survey designers have a number of options for testing their survey instrument to make sure respondents interpret the questions as intended and are not struggling to answer any questions. During the survey development phase, designers could conduct focus groups or cognitive interviews in which researchers get the chance to talk with respondents to better understand how they interpret the questions. Surveys that are final, or near final, should be pre-tested with a small number of respondents during which the researcher monitors the interviews.

Focus groups are typically used in the early stages of questionnaire design when researchers need to get additional information to determine the appropriate subject matter of a survey, response options, or test trial question wording. Because they are conducted in a group setting with multiple people at once, researchers can gain the opinions of a number of people in a short time.

Cognitive interviewing is done once survey questions are crafted and often the instrument is still in draft form but closer to final. Cognitive interviewing is a technique in which the researcher reads the question and the respondent provides an answer as usual. But after the response is given, the researcher discusses the meaning of the question with the respondent. The researcher typically asks the respondent to give his interpretation of the question and key question words. Survey researchers may discover that respondents may interpret a question very differently than what the researcher intended. Cognitive interviews are done one-on-one so that it takes time to gain information from multiple respondents.

Once a survey is ready to be fielded, researchers should conduct a small number of interviews in which the survey designer listens to the actual interviews while they are being conducted. Such monitoring is one of the only ways a survey designer can hear the full interview from the respondent's perspective. The designer will hear if respondents struggles to understand questions, have difficulty providing an answer that fits the response options, if the interview is too long or repetitive and respondents become impatient compromising data quality.

All of these techniques increase the costs and lengthens the time it takes to develop and field a survey. However, testing can prevent unpleasant surprises during data analysis if inconsistent and unreliable results come back. An inexpensive and quick survey is not a bargain if much of the data cannot be reported.

Conclusions

Survey research is a well-developed professional field in which a great deal of research has been conducted on how to collect the most reliable and valid data during an interview process. This research is vast and sometimes produces mixed results. It takes time and multiple studies to settle on a best practice. Evaluators and program implementers who do not have backgrounds in survey research do their best when designing and evaluating surveys. However, most have neither the time nor the background to make use of this research and apply it to their own work.

In this paper, we presented a number of best practices in survey design that are the result of years of research and experience. Evaluators can use this information to help them write their next survey while utility managers can use it to help them with their next survey review. Surveys play such a critical role in the evaluation of utility energy efficiency programs. It is important that we utilize the information coming from the world of survey research to ensure that we are collecting the best possible information on which to evaluate utility programs.

References

Bishop, G. 1987. "Experiments with the Middle Response Alternative in Survey Questions." *Public Opinion Quarterly* 51: 220-232.

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

Green, P. E., and V. R. Rao. 1970. "Rating Scales and Information Recovery – How Many Scales and Response Categories to Use?" *Journal of Marketing* 34: 33-39.

Krosnick, J. A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5: 213-236.

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

Krosnick, J. A., and S. Presser. 2010. "Question and Questionnaire Design." In J. D. Wright and P. V. Marsden (Eds), *Handbook of Survey Research (2nd Edition)*. Bingley, United Kingdom: Emerald Group Publishing.

Krosnick, J. A. et al. 2002. "The Impact of 'No Opinion' Response Options on Data Quality: Non-Attitude Reduction or Invitation to Satisfice?" *Public Opinion Quarterly* 66: 371-403.

Lissitz, R.W. and S. B. Green. 1975. "Effect of the Number of Scale Points on Reliability: A Monte Carlo Approach." *Journal of Applied Psychology* 60: 10-13.

Malhotra, N., J. A. Krosnick, and R. K. Thomas. 2009. "Optimal Design of Branching Questions to Measure Bipolar Constructs." *Public Opinion Quarterly* 73: 304-324.

Mondak, J. and B. C. Davis. 2001. "Asked and Answered: Knowledge Levels When We Will Not Take 'Don't Know' for an Answer." *Political Behavior* 23: 199-224.

Schuman, H. and S. Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context.* New York: Academic Press.

Schwartz et al. 1985. "Response Scales: Effects of Category Range on Reported Behavior and Subsequent Judgments." *Public Opinion Quarterly* 49: 388-395.

Simon, H.A. 1957. *Models of Man*. New York: Wiley.