The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating the Net-To-Gross Ratio

Richard Ridge, Ridge & Associates, Alameda, CA
Phillipus Willems, PWP Inc, North Potomac, MD
Jennifer Fagan, Itron, Inc., Madison, WI
Katherine Randazzo, KVD Research Consulting, El Cajon, CA

ABSTRACT

The net-to-gross ratio is one of the key parameters necessary to estimate a program's net energy and demand impacts and a variety of methods have been developed to estimate this critical parameter. Which technique one chooses depends of a number of factors such as time, money, data availability, and effect size. One non-experimental approach (i.e., one that does not involve any comparison group), the self-report approach (SRA), has been in use for at least 30 years and was developed in response to a number methodological challenges and policy-related issues. Because the SRA does not involve any formal comparison groups, it has been criticized as inherently biased and unreliable.

Professionals on both sides of this debate almost always fail to understand and appreciate the SRA's place within the larger evaluation framework, its evolving use over the last 30 years in the evaluation of education, mental health, juvenile justice, and energy efficiency programs and the resulting improvements in both its internal validity and reliability. To address this failure, this paper will discuss the rich history and evolution of this non-experimental approach in the broader evaluation community, and its more specific application in the evaluation of energy efficiency programs in California. This paper will also respond to some of the more common criticisms of the California SRA (the CA-SRA).

Background

A core feature of the use of any method to "determine" the impact of a program is the assumption that evaluation efforts have established a causal connection between the program and customer behavior. The establishment of causality is at the core of arguments for and against research methods aimed at establishing program impacts. When we couch our arguments in terms of causality, we benefit from being aware of the rich history of philosophical and methodological thinking and writing about causality that has taken place over the last 30 or 40 years. This awareness allows us to appreciate the issues and their complexity. It is important to establish that we in the energy efficiency field are not the first to address these issues and we can benefit from those who went before us. We therefore spend a few paragraphs of this paper rooting our arguments in this literature.

Historically, the social sciences have been heavily influenced by the *positivist* (or empiricist) philosophical approach to causality (Mohr 1995). Positivism asserts that causal relationships are not directly observable and therefore, causality is a matter only of observed regularities in associations of events (Hume 1737; Salmon 1998). According to this view, systematic, quantitative comparisons of events that produce correlations between variables are as close as researchers can get to causal explanation. This approach has sometimes been referred to as *variance theory* by Mohr (1995) and manipulation theory by Yang (2009) and emphasizes variables and the correlations among them. Variance theory is closely associated with statistical testing of hypotheses and, in particular, the regression model.

The randomized controlled trial (RCT), sometimes called the gold standard of the empiricist approach, consists of random assignment of subjects to two or more groups, only one of which receives

the treatment. One version of the RCT, the posttest-only control group design, is illustrated below in Figure 1 (Campbell and Stanley 1966). In Figure 1, R represents random assignment, *X* represents the treatment, O_1 represents the posttest for some variable *Y* for the treatment group and O_2 represents the posttest for the same variable for the control group.

R X O_1

R O_2

Figure 1. Posttest-Only Control Group Design

The comparison group represents what would have happened in the absence of the treatment. Any differences between the treatment and control group based on the posttests can be attributed to the treatment since all the other possible explanations have been effectively controlled for via randomization. That is, all the threats to internal validity have been addressed.[1] In other words, did in fact the experimental treatments make a difference in this specific experimental instance, i.e., if not *X* then not *Y*.

Because random assignment is not always possible, researchers have found it necessary to develop a number of other less powerful designs that are referred to as quasi-experimental. In such designs, researchers work with in-tact groups in natural social settings. However, such designs fail to control for all of the threats to internal validity (Campbell & Stanley 1966; Shadish, Cook, & Campbell 2002), particularly self-selection, making *definite* proof of causality impossible.

The Realist Perspective

The *realist* view of causal explanation represents a recent philosophical shift in the social sciences (Sayer 1984; Salmon 1998; Maxwell 2004). Realism defines causality as consisting not of regularities but of real (and in principle observable) causal mechanisms and processes, which may or may not produce regularities. It deals with events and the processes that connect them and analyzes relationships between events using data that retain relevant chronological and contextual connections. This perspective ascribes value to contextual factors and mental processes downplayed by the positivist approach to research. The realists view asserts that sometimes the reasons that people give for their behavior are indeed the causes of their behavior.

Mohr (1995) examines the central argument of the realists that stated reasons can be the causes of intentional behavior, often the focus in program evaluation. When people behave they do it for reasons that they can "observe" and when interrogated by a researcher they can report the most important reasons[2]. Of course, as he points out, at any given time we have good reasons for doing a lot of different things, but we do not always act on all of these reasons. We do only one thing, perhaps for only one of the many reasons. The other reasons might not have caused anything. Here, he introduces the concept of the "operative reason" which he defines as: ". . . the reason that actually operates to

[1] The traditional threats to internal validity identified by Campbell and Stanley (1966) are: 1) history, 2) maturation, 3) testing, 4) instrumentation, 5) regression, 6) selection, 7) mortality, and 8) interaction of selection and maturation, etc. However, as some have pointed out, even these designs are not perfect due to differential attrition (experimental mortality) from the experimental groups and the constraints (Shadish, Cook, and Campbell 2002) on external validity (Bickman & Reich 2009).

[2] Mohr (1995) provides a more detailed analysis of *physical* causality and *factual* causality, with the former associated with the realist position. Unfortunately, space does not permit a full exegesis of this distinction.

produce the behavior performed – and to stipulate that the operative reason is different from all the others in that it was the strongest" (265). That is, people undertake a particular action in order to achieve a certain goal because a particular reason was the strongest among all the competing reasons for undertaking that action.

This distinction between the empiricist and realist schools applies not only to the social sciences but also to the natural sciences, separating more experimental fields like physics and chemistry from those that deal with relatively unique situations, including evolutionary biology and geology. Gould (1989) observed:

> Historical science is not worse, more restricted, or less capable of achieving firm conclusions because experiment, prediction, and subsumption under invariant laws of nature do not represent its usual working methods. The sciences of history use a different mode of explanation, rooted in the comparative and observational richness of our data. (279)

We might add to this list of sciences that do not rely on experimental methods evolutionary biology, geology, and paleontology. This realist view of causation is compatible with and supports all the essential characteristics of qualitative research:

- If causal processes can be directly observed, then this supports the emphasis placed by many qualitative researchers on directly observing and interpreting social and psychological processes. It is possible to observe causal process in single cases without requiring comparison of situations in which the presumed cause is present or absent.
- Seeing context as intrinsically involved in causal processes supports the insistence of qualitative researchers on the explanatory importance of context.
- That mental events and processes are real phenomena that can be causes of behavior supports the fundamental role that qualitative researchers assign to meaning and intention in explaining social phenomena and the essentially interpretative nature of our understanding of these.
- In claiming that causal explanation does not inherently depend on pre-established comparison, it legitimizes qualitative researchers' use of flexible and inductive designs and methods.

This distinction between positivist/empiricist and realist approach is very similar to a distinction developed by Mohr (1982; 1995, 1996) between *variance theory* and *process theory*. Variance theory deals with variables and correlation, quantitative measurement, and experimental or correlational designs. Process theory deals with events and the processes that connect them; it is based on an analysis of the causal processes by which some events influence others. Maxwell (2004) also notes that:

> Process theory is not merely 'descriptive,' as opposed to 'explanatory' variance theory; it is a different approach to explanation. Experimental and survey methods typically involve a 'black box' approach to the problem of causality; lacking direct information about social and cognitive processes, they must attempt to correlate differences in output. Qualitative methods, on the other hand, can often directly investigate these causal processes. (p. 249)

However, as Maxwell (2004) points out, both the variance theory and the process theory face potential validity threats and each has its particular strengths.

The realist central argument is that qualitative research methods are as valid in determining causal explanation as purely quantitative ones, if they are well-designed to avoid threats to the validity of explanation. There are various research strategies that can be applied to qualitative research to best address these potential threats. The strategies are particularly productive when used in concert with a detailed theory of explanation in a given case, which can inform research design and interpretation of evidence as well as aid in developing alternative theories to be ruled out. Strategies typically associated

with variance theory, strategies of intervention and comparison, observation and analysis, strategies to develop and assess all alternative explanations, and triangulation can all provide important checks on bias and uncertainty in qualitative research. Four such strategies are described below.

- Intervention and comparison, commonly associated with quantitative research methods, are certainly compatible with qualitative research. Mixed-method research makes particular use of intervention to provide a detailed account of the process by which a particular statistical regularity occurs and to confirm statistical conclusions. Comparison without a formal control group but across sites or cases is also very helpful in identifying causal influences by providing evidence of altered or absent presumed influences. Comparisons can also be made between those in the treatment group and typical behavior (i.e., standard practice) within the same industry.
- Strategies can also be applied to the observation and analysis of causation that can help to address threats to validity. These strategies are unique to qualitative research, because they rely on observing processes rather than just end results. Such observations can also provide increased opportunities for developing and eliminating alternate hypotheses. Such "rich data" (also known as "thick description") give a more complete picture of context and process and therefore help to "counter the twin dangers of respondent duplicity and observer bias" by increasing the amount and varying the type of information necessary to confirm or deny a theory.
- Strategies for developing and assessing rival hypotheses can also guard against such threats. The *modus operandi* approach (Scriven, 1995) to alternative explanations is the most simple, relying on the researcher to identify other hypotheses and search for evidence of them in the data. This method is clearly vulnerable to researcher bias, as it may be difficult for researchers to identify other competitive theories and to detail them sufficiently to be able to test them. A similar strategy is that of identifying discrepant evidence and negative cases, which challenge the prevalent theory, explaining it away or adapting the hypothesis to incorporate the conflicting evidence.
- Triangulation, using a variety of research methods and data sources, is a strategy adopted before the data are collected and reduces the risk of systematic biases. However, there is also a risk that triangulation can make methodological or research bias harder to detect, by employing methods or data sources that are vulnerable to the same biases. Using the feedback of research participants to check conclusions and methods, can aid in identifying biases and alternative explanations, but subjects' inputs are subject to biases and influences of their own, not least of which is the need to reach consensus with the researcher.[3] Ultimately, Maxwell (2004) argues, ". . . validity threats are ruled out by evidence, not methods; methods need to be selected for their potential for producing evidence that will adequately assess these threats" (259)

In summary, qualitative and quantitative research methods each have their own threats to internal validity, as well as recourse to research design and implementation strategies to address these threats. Common to both methods is the danger of oversimplification of causal processes. Out of context, simple statements of linear relationships may be more distorting than illuminating. Qualitative analysis has an important advantage over quantitative in this respect: by providing insight into the composition of those relationships, it can aid in the strategic design of future research and help to advance practical action in the field studied.

[3] See discussion below of socially desirable responses and various strategies that have been developed to minimize this bias.

The Program Evaluation Framework

The development of the realist position paralleled the recognition on the part of many evaluators that the evaluation designs typically associated with the positivist approach were not always possible (Weiss 1972; Weiss & Rein 1972). As a result, many evaluators began to explore alternatives that would allow them to generate causal conclusions (Guba & Lincoln 1981; Cronbach 1982). This approach, consistent with the realist view, argues that qualitative research methods are as valid in determining causal explanation as purely quantitative ones, if they are well-designed to avoid threats to the validity of explanation.

The hallmark of scientific inquiry is the ability to eliminate alternative explanations and contradictory evidence. Research, whether quantitative or qualitative, must be meticulously designed to identify, detail, and test rival hypotheses. The *modus operandi*, introduced earlier, which, Mohr (1996) argued, shows a "distinct basis of efficacy" with respect to causation. Imagine that an outcome has occurred and the task of the evaluator is to demonstrate that the program, A, has caused the observed outcome, B. There are several other possible causes of B, such as C, D, and E. Each of these has a " signature" that has been defined by Mohr (1994) as: either or both (a) a mechanism or a known causal chain of events by which A, C, D, and E would lead to B and (b) the occurrence of other events in addition to A that are logically associated with or attributable to an active B, or C, or D, or E. The task of the analyst is to show that the signature of A has indeed been actualized, whereas the signature of each of the other possible or plausible causes has not. This basis of determining causality is relied upon heavily in many areas, such as detective work, cause-of death determination, medical diagnosis, and troubleshooting in connection with machinery, as in auto repairs (see *Car Talk*, National Public Radio). Chen (1990) and Rogers (2000) argues that sound program theories and logic models can provide valuable assistance in identifying the plausible set of rival hypotheses.

Most recently, Scriven (2009) has perhaps been one of the most outspoken champions of a process theory approach to causation. Scriven argues that the foundation of what constitutes cause in both science and law is based on the notion that causation is directly and reliably observable in everyday life. He further observes that:

> . . . good scientists have been entranced by the paragon of experimental designs, the randomly controlled trial or RCT, and illicitly generalized this into the required standard for all good causal investigation. It is suggested here that this view is completely refuted by a careful look at the way astronomy, epidemiology, engineering, geology, field biology, and many other sciences establish a causal conclusions to the highest standards of scientific (and legal) credibility. (151)

He even quotes Cook and Campbell to good effect: ". . . we do not find it useful to assert that causes are 'unreal' and are only inferences drawn by humans from observations that do not themselves directly demonstrate causation" (140).

Other evaluators have developed other approaches to demonstrating causality using non-experimental methods. For example, Yin (1994) provides guidelines for assessing causal relationships using case studies. Tashakkori and Teddlie (1998) also argue that their ". . . conceptualization of internal validity is not limited to experimental studies and causal relationships (p. 67)." Finally, consider Weiss (1997, 2000) who suggests that a theory-driven evaluation can substitute for classical experimental study using random assignment. She suggests that if predicted steps between an activity and an outcome can be confirmed in implementation, this matching of the theory to observed outcomes will lend a strong argument for causality: "If the evaluation can show a series of micro-steps that lead from inputs to outcomes, then causal attribution for all practical purposes seems to be within reach" (Weiss 1997, 43).

Some, such as Tashakkori and Teddlie (1998), recommend a mixed methodology in which quantitative and qualitative approaches are combined to improve internal validity. Or, consider Patton

who concludes that: "The field has come to recognize that, where possible, using multiple methods – both quantitative and qualitative – can be valuable since each has its strengths and one approach can often overcome weaknesses of the other (p. 267)." Applying these ideas to the California energy efficiency evaluation context, the 2005 Protocols provide that the use of both experimental/quasi-experimental and non-experimental methods are available to evaluators at the enhanced level of rigor in the 2005 Protocols.

The Energy Efficiency Program Evaluation Framework

As mentioned earlier, evaluators of energy efficiency programs also recognized, like evaluators in other fields, that there are situations in which the standard quantitative approaches involving comparison groups are not always possible. For example, in the industrial sector, three barriers are immediately apparent. First, there is an expected very small signal to noise ratio (low statistical power) in a participant/nonparticipant billing analysis i.e., the expected difference in monthly energy use between participants and nonparticipants was too small to detect reliably compared to other sources of variation in kWh that vary greatly across individual industrial sites. In addition, large industrial customers targeted by the program have been contaminated by participation in energy efficiency programs in prior years making it very difficult to find true nonparticipants. Finally, even if the first two problems were absent, the large industrial customers targeted by the program are each unique making it unlikely that one could find a group of nonparticipants that could be matched with participants on critical variables.

Also, consider new construction programs which also eventually confronted the problem that many of the large residential and nonresidential developers, architects and engineering firms had also been contaminated by participation in energy efficiency programs in prior years making it very difficult to find true nonparticipants. Over the years, a wide variety of methods such as discrete choice, difference-of-differences, and econometric modeling, had been used to estimate the net energy and demand impacts of these programs. However, these approaches have become, over time, increasingly unreliable and have produced implausible results. Thus, Chappel et al. (2005) concluded that, based on an assessment of these methodologies and a review of the needs of the 2004-05 Building Efficiency Assessment (BEA) study, the self-report approach is the most appropriate one for evaluating this complex and diverse program and market.

Again, finding a true nonparticipant who has not been contaminated by exposure to some energy efficiency treatment has become very difficult making it very difficult to model the effects of a single intervention using a variety of statistical techniques that relied on the comparison of participant and nonparticipant data.

Of course, there are also other budgetary or timing constraints that might prohibit the use of the quantitative approaches. The expected magnitude of the savings for a given program might not warrant the investment in a perhaps more expensive evaluation design that could involve a billing analysis or a discrete choice analysis of both participants and nonparticipants (assuming such group is even available). Or, key stakeholders might not want to wait for a billing analysis, which typically requires up to 12 months of post-implementation consumption data, to be completed. And with a small signal to noise ratio, the sample sizes necessary for the required statistical power can be prohibitively expensive.

In 1993, the recognition that methods involving comparison groups were not always feasible was first formalized in the energy efficiency field in California in the *Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management (DSM) Programs* (1993 Protocols). Based on this recognition, the SRA (hereafter referred to as the California SRA (CA-SRA))

was allowed as one way to estimate the NTGR[4] (a measure of the strength of the causal relationship between the program and the decision to install energy efficient equipment). From 1994 through 1998, the IOUs were responsible for conducting process and impact evaluations, guided by the 1993 Protocols[5], of their energy efficiency programs. During this time period, all IOUs used the CA-SRA along with other techniques as approved by the CPUC.[6] It is important to note that, although not required by the 1993 Protocols, the IOUs often employed triangulation (the use of two or more techniques to increase accuracy) in estimating net impacts. They did this because they, as well as the CPUC, understood that there is error associated with any single method, that the results would be used in integrated resource planning (IRP), and that the results would be used in a very direct way to determine their earnings on their energy efficiency investments.

Beginning in 2006, the Energy Division of the California Public Utilities Commission assumed responsibility for conducting all impact evaluations. In 2005, the *California Energy Efficiency Evaluation Protocols: Technical, Methodological and Reporting Requirements for Evaluation Professionals* (2005 Protocols) were developed to guide the evaluation of the programs in the 2006-08 funding cycle. Again, the CA-SRA was permitted with the agreement of the IOUs. However, the 2005 Protocols *explicitly* require (for the same three reasons listed above) triangulation for programs assigned the *enhanced* level of evaluation rigor[7].

This emphasis on triangulation throughout the last 15 years reflects a desire that the situations in which one was forced to rely solely on any one method would be rare. Of course, whether one is able to use one or more than one method, there will still be uncertainty surrounding the savings estimates owing to such things as sample error, measurement error, and the failure of multiple methods to arrive at the same answer. If the earnings mechanism established by regulators can either penalize or reward a utility if the savings vary by as little as 1 percent either way, then the pressure on evaluators to produce perfectly accurate and precise estimates of savings would be enormous. What policy makers and regulators sometimes forget is that no measurement system, no matter how rigorous, within the broader evaluation community, can meet that standard of accuracy. When evaluators fail to deliver the level of accuracy and precision required by regulators, one should not flog the evaluators and condemn their evaluation methods. Rather, one should change the regulatory framework from the high-stakes system of rewards and penalties so that they are more consistent with best evaluation practices.

The CA-SRA

Space limitations allow only a brief description of the CA-SRA and its relationship to the literature described earlier. The CA-SRA, rooted in the realist tradition, is a non-experimental approach that produces an estimate of the NTGR, an index of program influence. The NTGR is required by the 2005 Protocols and is used to adjust estimated gross energy and demand impacts in order to produce net energy and demand impacts, i.e., those impacts attributable to the program.

---

[4] The NTGR typically varies from 0 to 1 and represents the proportion of the gross savings that are attributable to the program.

[5] Appendix J (*Quality Assurance Guidelines For Statistical, Engineering, and Self-Report Methods for Estimating DSM Program Impacts*) to the 1993 Protocols provided a listing and discussion of the essential issues that should be considered by evaluators using self-report methods, together with some recommendations on reporting the strategies used to address each issue.

[6] Two additional methods, both of which involved the use of a comparison (not control) group, were discrete-choice analysis and billing analysis.

[7] Tashakkori and Teddlie (1998) classify such triangulation as *parallel mixed analysis,* which, they argue, is probably the most widely used mixed data analysis strategy in the social and behavioral sciences.

The CA-SRA involves asking one or more key participant decision-makers a series of closed and open-ended questions about their motivations for installing the efficiency equipment, about whether they would have installed the same EE equipment in the absence of the program, to establish the temporal precedence of the program, as well as questions that attempt to rule out rival explanations for the installation (Weiss 1972; Scriven 1976; Shadish 1991; Wholey et al. 1994; Yin 1994; Mohr 1995; Rogers et al. 2000; Donaldson, Christie, & Mark 2008). In the simplest case (e.g., residential customers), the CA-SRA is based primarily on quantitative data while in more complex cases in the nonresidential programs the CA-SRA is strengthened by the inclusion of additional quantitative and qualitative data which can include, among others, in-depth, open-ended interviews, direct observation, and review of customer and program records [8]. Many evaluators believe that additional *qualitative* data regarding the economics of the customer's decision and the decision process itself can be very useful in supporting or modifying *quantitatively*-based results (Britan, 1978; Weiss and Rein, 1972; Patton, 1987; Tashakkori and Teddlie, 1998; Cook, 2000). In early 2007, the Energy Division published the *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches* which contained 17 recommendations for further improving the validity and reliability of the CA-SRA (available at www.calmac.org) [9]. The output of the CA-SRA is an index (NTGR), a single number representing program influence. An estimated NTGR is required by the CPUC to adjust estimated gross impacts[10].

In 2007, the CPUC formed two groups (the Residential and Non-Residential NTGR Working Groups) comprised of nationally recognized experts in the use of the SRA to consolidate the lessons learned over the last 15 years in order to make further improvements in the CA-SRA. The primary objectives of this work were to address what appeared to be a systematic overestimation of freeridership (Ridge, 2001) and to produce standardized questionnaires, methods, and algorithms that could be used by all contract groups evaluating the energy efficiency programs in California for 2006-08. A more detailed description of the residential and non-residential CA-SRAs can be found at www.calmac.org.

We conclude this section by stressing that it does not make sense to paint all efforts to estimate NTGRs using the self-report approach, one version of which is the CA-SRA, with the same brush as some critics have done. Distinctions must be made between those efforts that conform to best practices in the use of this technique such as the CA-SRA and those that don't.

A Response to Critics

Over the years, a number of criticisms and arguments (many of which are of the straw man variety) have been leveled at the CA-SRA. We will very briefly address those that seem most important.

[8] Of course, even in the simplest cases, an evaluator is free to supplement the analysis with additional quantitative and qualitative data such as interviews with architects and engineers involved in residential new construction or HVAC installers and a review of available market share data.

[9] In 2003, PA Government Services prepared another comprehensive, although different set of comprehensive guidelines for the use of the self-report approach. The report, "Standardized Methods for Free-Ridership and Spillover Evaluation – Task 5 Final Report (Revised)" was prepared for National Grid, NSTAR Electric, Northeast Utilities, Unitil, and Cape Light Compact.

[10] Since 2004, the CPUC only adjusts for freeridership, rather than all contributions to net impacts such as participant and nonparticipant spillover.

Legitimacy

Some have leveled the general criticism that the CA-SRA is not a legitimate social science tool for establishing causality. This paper, it is hoped, has made some progress in refuting this criticism.

Turbulent Environment

In any evaluation, as the number of alternative hypotheses grows, the task of teasing out the effects of a single intervention becomes more challenging, i.e., a large portion of the population has been contaminated by other energy efficiency interventions or events in the marketplace such as Energy Star, Flex Your Power, efforts by such retailers as Wal Mart, other PGC-funded programs and growing awareness of the dangers posed by global warming. This is the case whether one is using quasi-experimental or one of the various realist approaches to causality. To argue that the increasing number of energy efficiency interventions makes it *impossible* to assess the efficacy of any given program, one must show that this environment is more challenging than that faced by evaluators in other arenas such as education, mental health, and advertising. We see no compelling evidence that this is the case. In fact, in these other turbulent environments, evaluators continue to evaluate a wide variety of interventions using a growing number of innovative techniques and designs in order to inform important decisions.

Nonlinear Approach

Peters and McRae (2009) argue that the CA-SRA is based on a false assumption, that the route by which the program reaches the energy user is linear (" . . . the participant seeks the solution to a problem or to purchase a piece of equipment, learns of the efficiency opportunity promoted by the program, and decides to take the efficient action."). In fact, the CA-SRA explicitly recognizes that the route is nonlinear by attempting to identify the engineers, architects, vendors etc. who were most important in the customer's decision to participate and uncover the various ways in which the utility programs might have influenced these market actors. In the turbulent energy efficiency environment, the CA-SRA is focused on identifying the multiple lines of influence over time and recognizes that out of context, simple statements of linear relationships are more distorting than illuminating (Rogers 2000).

Recall

One of the problems inherent in the CA-SRA is that we are asking customers to recall what has happened in the past. It is well known in the interview literature that the more factual and concrete the information the survey requests, the more accurate responses are likely to be. Where we are asking for motivations and processes in situations that occurred one or two years ago, there is room for bias. In order to minimize the problem of recall, CA-SRA interviews should be conducted with the decision maker(s) as soon after the installation of equipment as possible (Stone et al. 2000).

Subjective

Various stakeholders have criticized the CA-SRA as being too subjective. The CA-SRA collects a variety of qualitative and quantitative evidence (e.g., corporate documents, past purchase patterns, closed-ended questions regarding motives and their strength from multiple sources). If an evaluator is able to provide some *evidence* for the mechanisms involved in the hypothesized causal links then this is not merely subjective.

Treating Ordinal Data as Interval

We begin by distinguishing ordinal data from interval data. Ordinal data have order, but the interval between measurements is not meaningful (i.e., moving from a 1 to a 2 is not necessarily the same as moving from a 5 to a 6). As a result, it is technically inappropriate to apply basic mathematical operations such as the calculation of means. On the other hand, interval data have meaningful intervals between measurements which support a variety of arithmetic operations. In the CA-SRA, what we are attempting to measure, among other things, is a participant's perception of the influence of the utility program on their decision to implement the energy-efficient measure. Because this is not something that is directly observable and measureable, we must rely on answers to a series of questions regarding the reasons for the installation. To assess the strength of any reason, we have chosen response categories along a 0-10 scale since the strength of the reasons cannot be adequately captured by a "yes" or "no" response. Making this choice means that we are not certain that moving from a 1 to a 2 is the same increase as moving from a 5 to a 6. However, we are willing to treat the responses as "sufficiently" interval for our purposes. While the debate over treating ordinal scales as interval has been going on for some time, there is strong support in the social science literature that treating ordinal variables as interval yields results that are both *meaningful* and *useful* to decision makers (Velleman & Wilkinson 1993; Tashakkori & Teddlie 1998)[11]. The ordinality of the observed data presumably reflects an underlying interval scale that just can't be measured at that level. Therefore, the lack of consistency in the distance between measured levels (1-2 versus 5-6) constitutes measurement error, something researchers of all kinds live with constantly. There is no reason to think that this measurement is not randomly distributed, and, therefore, of the most benign kind. At the very least, there is no reason to assume that the varying size of intervals biases responses upward or downward.

The Meaning and Calculation of the NTGR

Some have demonstrated that the calculation of the core NTGR can vary dramatically depending on the algorithm and the weights that an evaluator assigns to the different components. That changing the algorithms or weights results in big changes is obvious. It is equally obvious that algorithms and weights must be developed experienced professionals who understand that algorithms and weights have to be transparent, plausible and defensible and that they must be subjected to thoughtful sensitivity analysis. Many of the tools in science can provide bad results when done poorly. That does not prove the tool is bad only that the scientist is doing bad work - whether it is in free-ridership estimation, regression analysis, DOE2 modeling, or other fields with regression, gas chromatography, blood lab testing, reading biopsies, etc.![12]

Socially Desirable Responses

Another commonly recognized motivation for biased answers is that some people will like to portray themselves in a positive light; *e.g.*, they might like to think that they would have installed energy-efficient equipment without any incentive (the socially desirable response). This type of motivation could result in an artificially low net-to-gross ratio. The existence of the socially desirable response has been a perennial problem for survey researchers. Critics (Peters & McRae, 2008) appear to think that simply leveling this criticism is sufficiently damning. Unfortunately, they appear unwilling to

[11] Note that measurement of variables such as intelligence, depression and quality of life are all ordinal but are usually interpreted as if they were interval.

[12] These observations, with which we wholeheartedly agree, were provided recently by Lori Megdal.

acknowledge the various methods and techniques (Bradburn, Sudman, & Wansink 2004; Lyberg et al. 1997; Groves et al. 2004) that have been developed to address this potential source of bias and the extent to which these have been incorporated into the CA-SRA. For example, Bradburn, Sudman, and Wansink (2004) provide a checklist of 13 techniques for minimizing this bias including using data from knowledgeable informants (e.g., vendors, installers, etc.), attempting to validate the answers, and using both closed and open questions. These three are among a number of techniques that have been incorporated into the CA-SRA. Of course, it is possible that a respondent might exaggerate the importance of the program because they want the program and its rebates to continue. Technically, this is not a case of the socially desirable response bias but does represent a type of biased response that should be mentioned. The same techniques used to reduce the socially desirable response bias can be used to mitigate this other type of bias.

Stated Intentions

Peters and McRae (2008) argue that asking the respondent what they would have done in the absence of the program (the so-called *counterfactual*) is fatally flawed, since people are notoriously bad about following through on their stated intentions. This indeed would be a fatal flaw if only a single counterfactual type of question were asked of a single decisionmaker. In fact, the approach in the residential and nonresidential sectors is far more robust. Some questions are designed to measure the counterfactual by asking the participant a number of questions about what they would have done in the absence of the program. However, other questions attempt to get at the *operative* reasons for installing the efficient equipment. As part of this set of questions, the respondent is prompted to consider program and other possible non-program influences that might have played a role in the decision. Still other questions attempt to establish the temporal precedence of the program (information and/or rebate), i.e., when the participant first heard about the program relative to their decision to install the efficient equipment. In the nonresidential sector, additional information is gathered from program files, vendor surveys, account representatives, interviews with industry experts, and other program documentation to construct an internally consistent story surrounding the decision to install the energy efficiency equipment.

Conclusions

We have demonstrated that the realist approach is a legitimate method with a firm grounding in the epistemological literature. A variety of qualitative evaluation methods have been developed over the last 30 years that are consistent with the realist approach. Such qualitative methods for assessing causality can be rigorous and even more so if they are combined with quantitative methods, and vice versa. Within the evaluation community, many leading experts have endorsed such an approach, although many do see it as complementing not supplanting experimental or quasi-experimental approaches, i.e., a version of the mixed method approach. The CA-SRA is consistent with these qualitative approaches developed in the broader evaluation community. In California, for projects with substantial savings that have been assigned the enhanced level of rigor, the 2005 Protocols require, and we agree, that two or more approaches of the available three (discrete choice with a comparison group, billing analysis with a comparison group, and the CA-SRA) must be used. Such a mixed method approach provides a much improved (not perfect) level of accuracy. For programs that have been assigned the standard or basic level of rigor and for which methods involving comparison groups are impossible, the CA-SRA can provide estimates of the NTGR that are sufficiently rigorous for determining the degree of program influence for assessing program efficacy.

We also argued forcefully that any regulatory set of rewards and penalties should never require a level of accuracy that exceeds the ability of any evaluators to provide. To do so places and unreasonable burden on evaluators and ensures a never ending, contentious and unproductive relationship among program implementers and the regulatory community.

Finally, the straw man criticisms that have been lodged against the CA-SRA must be recognized for what they are; accusing the CA-SRA of methodological sins it never committed then deploring its lack of virtue.

References

Bickman, Leonard and Stephanie M. Reich. 2009. Randomized Controlled Trials: A Gold Standard with Clay Feet? In Stuart I. Donaldson, Christina A. Christie, and Melvin Mark (Eds.). *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Los Angeles, CA: SAGE Publications.

Britan, G. M. 1978. "Experimental and Contextual Models of Program Evaluation." 1978. *Evaluation and Program Planning* 1: 229-234.

Campbell, Donald T. and Julian Stanley. 1963. *Experimental and Quasi-experimental Designs for Research*. Boston, MA: Houghton Mifflin Company.

Chen, Huey-Tsyh. 1990. *Theory-Driven Evaluations*. Newbury Park, CA: SAGE Publications.

Cronbach, L. J. 1982. *Designing Evaluation and Social Action Programs*. San Francisco: Jossey-Bass.

Donaldson, Stewart I, Christina A. Christie and Melvin Mark (Eds.) 2009. *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Los Angeles, CA: SAGE Publications.

Hume, David. 1737. *Enquiries Concerning the Human Understanding and Concerning The Principles of Morals*. Oxford University Press 2nd Ed. 1957.

Guba, E.V. and Y. S. Lincoln. 1981. *Effective Evaluation*. San Francisco: Jossey-Bass.

Guba, E. G. 1978. *Toward a Methodology of Naturalistic Inquiry in Educational Evaluation* (CSE Monographic Series in Evaluation No. 8). Los Angeles: Center for the Study of Evaluation.

Maxwell, Joseph A. 2004. "Using Qualitative Methods for Causal Explanations." *Field Methods*, Vol. 16, No. 3, 243-264

Mohr, Lawrence B. 1995. *Impact Analysis for Program Evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

PA Government Services. 2003. *Standardized Methods for Free-Ridership and Spillover Evaluation – Task 5 Final Report (Revised)*. Prepared for National Grid, NSTAR Electric, Northeast Utilities, Unitil, and Cape Light Compact.

Pacific Gas & Electric, San Diego Gas & Electric, Southern California Edison, Southern California Gas, California Energy Commission, Office of Ratepayer Advocates (CPUC) and the Natural Resources Defense Council. 1993. *Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management (DSM) Programs.* As adopted by the California Public Utilities Commission Decision 93-05-063.

Patton, Michael Quinn. 1987. *How to Use Qualitative Methods in Evaluation*. Newbury Park, California: SAGE Publications.

Peters, Jane and Marjorie McRae. 2008. Free-Ridership Measurement If Out of Sync with Program Logic . . . or, We've Got the Structure Built, but What's Its Foundations? In the *Proceedings of the 2008 ACEEE Summer Study on Energy Efficiency in Buildings*, ACEEE.

Ridge, Richard and Mike Rufo. 2001. *Improving the Standard Performance Contracting Program: An Examination of the Historical Evidence and Directions for the Future*. Prepared under contract to Xenergy and submitted to the Southern California Edison Company.

Ridge, Richard, Ken Keating, Lori Megdal, and Nick Hall. 2007. *Guidelines for Estimating Net-To-Gross Ratios Using the Self Report Approach*. Prepared for the California Public Utilities Commission.

Rogers, Patricia J., Timothy A. Hacsi, Anthony Petrosino, and Tracy A. Huebner (Eds.) 2000. *Program Theory in Evaluation: Challenges and Opportunities*. San Francisco, CA: Jossey-Bass Publishers.

Salmon, Wesley C. 1998. *Causality and Explanation*. New York: Oxford University Press.

Scriven, Michael. 1976. Maximizing the Power of Causal Explanations: The Modus Operandi Method. In G.V. Glass (Ed.), *Evaluation Studies Review Annual* (Vol. 1, pp.101-118). Beverly Hills, CA: Sage Publications.

Scriven, Michael. 2009. Demythologizing Causation and Evidence. In Stuart I. Donaldson, Christina A. Christie, and Melvin Mark (Eds.). *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Los Angeles, CA: SAGE Publications.

Shadish, Jr., William R. and Thomas D. Cook, and Laura C. Leviton. 1991. *Foundations of Program Evaluation*. Newbury Park, CA: Sage Publications, Inc.

Stone, Arthur A., Jaylan S. Turkkan, Christine A. Bachrach, Jared B. Jobe, Howard S. Kurtzman, and Virginia S. Cain. 2000. *The Science of the Self-Report: Implications for Research and Practice*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Tashakkori, Abbas and Charles Teddlie. 1998. *Mixed Methodology: Combining Qualitative and Quantitative Approaches*. Thousand Oaks, CA: SAGE Publications.

Weiss, R. S. and M.Rein. 1972. The Evaluation of Broad-Aim Programs: Difficulties in Experimental design and an Alternative. In C. H. Weiss (ed.) *Evaluating Action Programs: Readings in Social Action and Education*. Boston: Allyn and Bacon.

Weiss, Carol H. 1997.Theory-Based Evaluation: Past, Present, and Future. In Debra J. Rog and Deborah Fournier (eds.) *Progress and Future Directions in Evaluation: Perspectives on Theory, Practice, and Methods*. San Francisco: Jossey-Bass Publishers.

Weiss, Carol H. 1998. *Evaluation*. Upper Saddle River, New Jersey: Prentice Hall.

Wholey, Joseph S., Harry P. Hatry and Kathryn E. Newcomer. 1994. *Handbook of Practical Program Evaluation*. San Francisco, CA: Jossey-Bass, Inc.

Yin, Robert K. 1994. *Case Study Research: Design and Methods*. Thousand Oaks, CA: SAGE Publications.