Self-Report Methods for Estimating Net-to-Gross Ratios in California: Honest!

Richard Ridge, Ridge & Associates, Alameda, CA
Philippus Willems, PWP Inc, North Potomac, MD
Jennifer Fagan, Itron, Inc., Madison, WI

ABSTRACT

This paper demonstrates how the current approach to the calculation of net-to-gross ratios (NTGRs) for impact evaluations of nonresidential programs in California builds on established methods of social science research. As NTGR analysis takes on added importance in light of aggressive energy savings goals and evaluation-driven incentive mechanisms, energy service professionals may question the accuracy and legitimacy of self-report-based methods relative to quantitative approaches for estimating net impacts. This paper will show how the techniques being used in California are consistent with analytical methods that have been used in the social sciences for many years, and are both practical and theoretically sound.

At the center of the NTG approach used in California is the calculation of a NTGR based on a series of questions that attempt to assess: 1) the counterfactual, 2) the direct influence of the various program components, and 3) the overall influence of the program relative to other factors. In addition, in accordance with a set of protocols developed for the California evaluations, the nonresidential NTGR approach provides for the incorporation of data from other sources such as corporate procurement policies regarding the purchase of energy efficient measures within a given corporation and standard practice regarding the installation of energy efficient measures within a given industry. The calculation of a NTG ratio in this manner may be viewed in the same light

as the creation of other economic and social research indices that use self-reported data, such as those for consumer confidence or IT manager purchase intentions.

The Social Science Framework

Historically, the social sciences have been heavily influenced by the positivist (or empiricist) philosophical approach to causality. Positivism asserts that causal relationships are not directly observable (Hume, 1737), and therefore defines causality as a matter only of observed regularities in associations of events. According to this view, systematic, *quantitative* comparisons of events that produce correlations between variables are as close as researchers can get to causal explanation[1]. Consequently, *qualitative* research methods have often been unfairly ignored or deemed speculative or amateur by researchers in the social sciences.

However, in the early 1970s, many evaluators began to realize that the evaluation designs typically associated with the positivist approach were not always possible (Weiss, 1972; Weiss and Rein, 1972). As a result, many evaluators began to explore alternatives that would allow them to generate causal conclusions (Guba and Lincoln, 1981; Cronbach, 1982). This approach, the so-called realist view, has emerged that 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 realist view of causal explanation represents a relatively recent philosophical shift in the social sciences. In this view, causality has been defined (Maxwell (1992, 12) as: ". . . as consisting not of regularities but of real (and in principle observable) causal mechanisms and processes, which may or may not produce regularities." This perspective ascribes value to contextual factors and mental processes downplayed by the positivist approach to research. Since it holds that causal processes are observable, it also provides strong support for case studies and other strategies that do not use strictly comparative experimental methods.

This divide between realism and empiricism underpins two differing approaches to research into causal explanation, whether in the social or natural sciences. Variance theory emphasizes variables and the correlations among them and is closely associated with statistical testing of hypotheses and in particular the regression model. Process theory, on the other hand, deals with events and the processes that connect them and analyzes relationships between events using data that retain the chronological and contextual connections. This distinction 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.

---

[1] Determining these causal connections was first referred to by Campbell and Stanley (1963) as internal validity, i.e., Did in fact the experimental stimulus make some significant difference in this specific instance?

Both types of causal explanation face potential validity threats, and each has its particular strengths. 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 Self-Report Method

Historically, like the evaluations of other social action programs, the standard approach to assessing causality was to conduct an experiment or quasi-experiment[2] in which data were collected from both participants and nonparticipants with the data being subjected to a variety of statistical analyses.[3] However, it soon became clear that there are situations that make this approach, for both technical and non-technical reasons, impossible. For example, in some cases, the expected magnitude of the savings, for a given *program* might not warrant the investment in an expensive evaluation design that could involve a billing analysis or a discrete choice analysis of both participants and nonparticipants. Or, key stakeholders might not want to wait for a sufficient amount of post-installation monthly consumption to accumulate so that a billing analysis to be completed. Also, if the relationship of the savings to the normal monthly variation in energy use is too small, then a billing analysis should not even be attempted owing to a lack of statistical power. This is one of the reasons why billing analysis is rarely, if ever, attempted in the industrial sector. Finally, in some cases, it might not be possible to identify a group of customers to serve as a comparison group since they have been exposed through prior participation or are in some other ways contaminated. So, for budgetary, timing, statistical, and research design issues, alternative approaches to the more traditional designs and analyses were developed.

Rooted in the realist approach to causation, the alternative approach to the evaluation of energy efficiency programs came to be known as the self-report approach (SRA) and deviates from the standard approach, described above, to assessing causality, i.e., internal validity. As will become clear, the term "self-report" as applied to the nonresidential sectors does not do justice to the methodological complexity of this approach which has deep roots in the realist wing of the evaluation community. To suggest that it only involves asking one key decisionmaker to hypothesize about what equipment they would have installed in the absence of the program is misleading.

Perhaps a better term can be applied to this approach such as the *program influence index* since the NTGR is nothing more than an index of the extent to which utilities have designed and implemented efficient program, i.e., programs that do not incent customers to do things that they would, absent the program, do anyway. Assessing the influence of the program by estimating the NTGR is not like measuring various attitudes for which many scales have been developed. In scaling, scores on items are

[2] In the literature, evaluations of energy efficiency and conservation programs that involve the use of a true experimental design are very rare.

[3] The use of such traditional evaluation designs are still possible and should be tried whenever possible in conjunction with non-experimental methods.

theoretically caused by the latent or underlying construct, i.e., they are "reflected" by the latent construct. With an index, scores on items (indicators) drive the total score of the index, i.e., the items/indicators "form" the constructed index score. DeVellis (1991; 9) points out that an index is ". . . made up of 'cause indicators' or items that determine the level of a construct."

Although still latent in many respects, formative items/indicators are not considered scales because their scores are not necessarily reflected by the latent construct. An often used example of formative items/indicators that result in an index is socioeconomic status (SES). Items or indicators might include income, education level, occupation, and dwelling type. Although some of these indicators have the latent property of not being directly observable, their scores are considered "forming" the index of SES and not vice versa. For example, SES is based partly on education. However, having more education is not caused by higher SES; it influences SES.

Developing an index of program influence is actually far more difficult than attempting to measure attitudes since causality in evaluation of energy efficiency programs has many dimensions involving, particularly in the nonresidential sector, the collection of information from more than just a single participant decision maker (e.g., vendors). It is similar in that we are trying to measure a latent construct called *program influence* that is not directly observable and therefore we must interpret the responses to various types questions put to participants and, in many cases, vendors combined with other information about company policy, standard practice etc. in order to assess the level of program influence.

In nonresidential cases, we are attempting to determine the extent to which the program influenced a participant to install an energy efficient measure(s). One minus this score is interpreted as freeridership. 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. Other questions attempt to get at the direct influence of the rebate and other forms of assistance on the decision to install efficient equipment. As part of this set of questions, the respondent is prompted to consider other possible non-program influences that might have played a role in the decision. Still other questions attempt to establish the chronology of when the participant first heard about the program and their decision to install the efficient equipment. These three different types of questions are trying to measure three slightly different things with some being more difficult than others for the respondent to assess. For example, it is easier for the respondent to recall whether they found out about the availability of the rebate before or after they decided to buy the efficient equipment than it is to imagine what they would have done in the absence of the program or assess the influence of the rebate. Nevertheless, all three types of questions provide information about the influence of the program that decision makers should find both meaningful and useful.

More specifically, the nonresidential SRA is a mixed method approach that involves asking one or more key participant decision-makers a series of structured and open-ended questions about their motivations for installing the efficiency equipment,

whether they would have installed the same EE equipment in the absence 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, and Mark, 2008). In the simplest case (e.g., residential customers), the SRA is based primarily on quantitative data while in more complex cases the 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 [4]. 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).

Challenges

Having presented a very brief history of these alternatives approaches and their motivation as well as the SRA itself, we move on to discuss a number of special challenges associated with the SRA that merit mentioning.

One of the problems inherent in asking program participants if they would have installed the same equipment or adopted the same energy-saving practices without the program is that we are asking them to recall what has happened in the past. Worse than that is the fact that what we are really asking them, among other things, is report on a hypothetical situation, what they would have done in the absence of the program. In many cases, the respondent may simply not know and/or cannot know what would have happened in the absence of the program. Even if the customer has some idea of what would have happened, there is, of necessity, uncertainty about it.

The situation just described is a circumstance ripe for invalid answers (low construct validity) and answers with low reliability, where reliability is defined as the likelihood that a respondent will give the same answer to the same question whenever or wherever it is asked. 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 hypothetical situations that occurred one or two years ago, there is room for bias (Stone et al., 2000). Bias in responses is commonly thought to stem from four origins.

Keep the program bias: First is the fact that some respondents may believe that claiming no impact for the program is likely to cause the program to cease, thus removing future financial opportunities from the respondent.

I'm a good/smart person bias: The second 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

[4] 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.

equipment without any incentive (the socially desirable response). This type of motivation could result in an artificially low net-to-gross ratio. Note that in the nonresidential sector, it is very possible that this bias might be offset or even go in the other direction since nonresidential customers are more likely to understand that giving credit to the program increases the likelihood that the program will continue to exist.

Intention to take action bias: The third hypothesized source of bias involves an interaction between the positive perception of taking energy efficiency actions, the often observed difference between stated intentions and actual behaviors, and the fact that the counterfactual outcome can not be viewed, by the participant or outsiders. One of the three main sections in the nonresidential NTGR survey involves a series of questions regarding the actions they would have taken if there had been no program. More specifically, this is asking the respondent to state their intentions with respect to purchasing the relevant equipment absent the program. Bias creeps in because people may intend many things that they do not eventually accomplish.

Complicated lines of influence bias: A fourth hypothesized source of bias arises when participants are asked to identify the reasons why they installed the energy efficient measures. Respondents might not always be able recall all the possible reasons and influences, the program information and rebate among them, and rank each in terms of its importance.

Beyond the fact that the situations of interest have occurred in the past and there are multiple sources of potential bias, the situations are often complex. No one set of questions can apply to all decision processes that result in a program-induced course of action. Some installations are simple, one-unit measures, while others involve many units, many different measures, and installations taking place over time. The decision to install may be made by one person or several people in a household, an individual serving as owner/operator of a small business, or, in the case of large commercial, industrial, or agricultural installations by multiple actors at multiple sites. Some measures may have been recommended by the utility for years before the actual installation took place, and others may have been recommended by consultants and/or vendors, making degree of program influence difficult to establish. Finally, some efficiency projects may involve reconfiguration of systems rather than simple installations of energy-efficient equipment.

Another factor that can complicate the SRA is that, in certain situations, the estimated NTGR combines (more often implicitly than explicitly) the probability of a decision/action occurring and whether the *quantity* of the equipment installed would have been the same. This can complicate the interpretation of the responses and the way in which to combine these types of questions in order to estimate the NTGR.

The SRA Guidelines

The *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches* was commissioned by the Energy Division of the California Public Utilities Commission (CPUC) to address the challenges list earlier with respect to reliability and validity[5]. For example, the Guidelines recommended interviewing all relevant decision makers shortly after the installation of the energy efficient equipment, building in consistency checks, using multiple questions, employing triangulation, ruling out of rival hypotheses, using a combination of quantitative and qualitative data, conducting sensitivity analysis, incorporating other documentation such as a company's procurement policies and standard practice in a particular industry, and, for situations when substantial savings are being claimed, using two analysts to independently review all data collected. The recommendations were meant to set the minimum acceptable standards for the use of the SRA to estimate net-to-gross ratios. The non-residential SRA has incorporated all of these recommendations.

Application of the Nonresidential Self Report Approach

For the analysis of the 2006-2008 California large commercial and industrial programs, a revised nonresidential self-report approach that incorporates the above considerations as reflected in the SRA guidelines is being used to assess net impacts. In recognition of the varying degrees of complexity of projects and the underlying decision processes, three levels of detail are being used – all built around the same **core** questions, but incorporating additional information sources and review as the size and complexity of projects increases. The most detailed, **Standard – Very Large Project** NTGR, is being applied to the 10-20% of all projects that are most complex and have the greatest expected gross savings. **Standard** NTGR with a less detailed level of analysis is applied to medium-sized projects, while **Basic** NTGR is applied to all remaining projects.

Sources of Information on Free Ridership

It is important to emphasize that despite the use of the term "self-report," there are in fact five sources of free-ridership information being utilized in this application of the self-report approach to the California commercial and industrial programs. Results from these five sources are cross-validated to ensure a consistent NTGR "story" emerges. If conflicting information is found in any one of the sources, the interviewers are required to circle back to those providing conflicting answers to provide a satisfactory explanation and to resolve the conflict.

[5] The type of validity that concerns us here is internal validity, which addresses the basic question: Is the observed change due to the program? The positivists and realists have each developed different methods for answering this question. The Guidelines are more consistent with the realist position.

These sources are described below.

1. **Program Files**. In addition to the essential data on measure cost, expected savings, and rebate; contact information; and vendor names, program files can contain information relevant to the net analysis, such as measure payback with and without the rebate.

2. **Decision-Maker Survey**. From the program files, program managers and site recruitment calls, we determine who made the decision to implement measures under the program. This individual is asked to complete a survey with structured questions regarding decision influences and likely actions in the absence of the program. There are three separate sets of questions that contribute to three components of the overall NTGR, integrated through the use of a NTGR scoring algorithm[6]. To determine the first component, participants are asked about the importance of program and non-program influences that may have led to their decision to implement the energy efficiency measure at the time that they did. These factors include:

   - the age or condition of the equipment,
   - information from a feasibility study or facility audit
   - the availability of an incentive or endorsement through the program
   - a recommendation from an equipment supplier, auditor or consulting engineer
   - previous experience with the program or measure
   - information from a program training course or marketing materials
   - availability of and competition for capital
   - a recommendation from program staff, a program vendor, or a utility representative
   - standard business practice
   - an internal business procedure or policy
   - concerns about global warming or the environment.

   In addition, the survey asks the decision maker to describe what equipment would have been implemented in the absence of the program, including the efficiency level and quantities of these alternative measures and the timing of the action. This information is used to account for deferred free ridership and to adjust the gross savings estimate for partial free ridership.

   Finally, the decision maker is asked to consider the most important program and non-program influences that have just been discussed and to rate the *relative* importance of each in their overall decision to implement the program qualifying measure.

   The decision maker survey contains a **core** set of questions that are used in the NTGR algorithm for **Basic** NTGR sites, and several supplemental questions for

[6] More details regarding the questionnaire and the algorithm are available upon request.

Standard NTGR sites (to help construct a "story" based on the information given). **Standard – Very Large Project** NTGR sites receive additional probing on various aspects of the decision making based on responses to specific questions. For example, if the respondent says a financial calculation was highly significant in their decision, they are asked about their financial criteria for investments and rationale for the current project. Similarly, if a corporate policy was a primary consideration, they are asked about the specific policy that led to adoption of the installed measure. These questions provide a deeper understanding of the decision making process and the likely level of program influence versus these internal policies and procedures.

3. **Vendor Survey**. A Vendor Survey is completed for all **Standard** and **Standard - Very Large** NTGR sites that utilized vendors, and for **Basic** NTGR sites that indicate a high level of vendor influence in the decision to implement the energy efficient measure. If respondents say the vendor was very influential in decision making, the vendor survey results will enter directly into the NTGR scoring. The vendor findings are also used to corroborate Decision Maker findings. Vendors are queried on the program's significance in their decision to recommend the program measures, and on their likelihood of recommending the same measure in the absence of the program.

4. **Other information**. For **Standard – Very Large Project** NTGR sites, secondary research of other pertinent data sources can provide valuable information on industry trends that help explain the context for a project decision. For example, this could include a review of standard and best practices through industry associations, industry experts, and information from secondary sources.

5. **Information from the Account Representative.** For nonresidential customers with large expected savings, utility account representatives provide similar background data, including the decision maker's name and contact information and the history of each project, including their perception of program or utility influence.

Table 1 below shows the data sources used in each of the three levels of free-ridership analysis. Although more than one level of analysis may share the same source (e.g., the decision maker interview), the amount of information utilized in the analysis may vary.

Table 1: Information Sources for Three Rigor Levels

NTGR Instruments	Program File	Account Rep Survey	Decision Maker Survey	Vendor Survey	Other Research Findings
Basic NTGR	√		√	√[1]	
Standard NTGR	√	√	√	√	
Standard NTGR – Very Large	√	√	√	√	√

[1] Only performed for sites that indicate a high level of Vendor influence, based on a score of 6 or greater.

Data Analysis and Integration

The analysis for the Basic NTGR is fairly mechanical, calculating a NTGR based on the answers to the closed-ended questions. However, the reliance of the Standard NTGR and Standard NTGR – Very Large on more information from so many different sources requires more of a case study level of effort. The SRA Guidelines point out that a case study is one method of assessing both quantitative and qualitative data in estimating a NTGR. A case study is an organized presentation of all these data available about a particular customer site with respect to all relevant aspects of the decision to install the efficient equipment. In such cases where multiple interviews are conducted eliciting both quantitative and qualitative data and a variety of program documentation has been collected, one will need to integrate all of this information into an internally consistent and coherent story that supports a specific NTGR.

Sometimes, *all* the quantitative and qualitative data will clearly point in the same direction while, in others, the *preponderance* of the data will point in the same direction. Other cases will be more ambiguous. In all cases, in order to maximize reliability, it is essential that more than one person be involved in analyzing the data. Each person must analyze the data separately and then compare and discuss the results. Important insights can emerge from the different ways in which two analysts look at the same set of data. Ultimately, differences must be resolved and a case made for a particular NTGR. Careful training of analysts in the systematic use of rules is essential to insure inter-rater reliability[7].

These rules should be as specific as possible and be strictly adhered to throughout the analysis. Such rules might include instructions regarding when the NTGR based on the quantitative data should be overridden based on qualitative data, how much qualitative data is needed to override the NTGR based on quantitative data, how to handle contradictory information provided by more than one person at a given site, how to handle situations when there is no decision-maker interview, when there is no appropriate decision-maker interview, or when there is critical missing data on the questionnaire, and how to incorporate qualitative information on deferred free-ridership.

One must recognize that it is difficult to anticipate all the situations that one may encounter during the analysis. As a result, one may refine existing rules or even develop new ones during the initial phase of the analysis. One must also recognize that it is difficult to develop algorithms that effectively integrate the quantitative and qualitative data. It is therefore necessary to use judgment in deciding how much weight to give to the quantitative versus qualitative data and how to integrate the two. The methodology and estimates, however, must contain methods to support the validity of the integration

[7] Inter-rater reliability is the extent to which two or more individuals (coders or raters) agree. Inter-rater reliability addresses the consistency of the implementation of a rating system.

methods through preponderance of evidence or other rules/procedures as discussed above.

For California, it was decided that supplemental data from non-core NTG questions collected through these surveys should be used in the following ways:

- Vendor interview data will be used at times in the direct calculation of the NTGR. It will also be used to provide context and confirming/contradictory information for Standard-Very Large decision maker interviews.

- Qualitative and quantitative information from other sources (e.g., industry data, vendor estimates of sales in no-program areas) may be used to alter core inputs only if contradictions are found with the core survey responses. Since judgments will have to be made in deciding which information is more compelling when there are contradictions, supplemental data are reviewed independently by two senior analysts, who then summarize their findings and recommendations and together reach a final NTGR value.

- Responses will also be used to construct a NTGR "story" around the project; that is they will help to provide the context and rationale for the project. This is particularly valuable in helping to provide guidance to program design for future years. It may be, for example, that responses to the core questions yield a high NTGR for a project, but additional information sources strongly suggest that the program qualifying technology has since become standard practice for the firm or industry, so that free ridership rates in future years are likely to be higher if program rules are not changed.

Findings from other non-core NTGR questions should also be used to cross-check the consistency of responses to core NTG questions. When an inconsistency is found, it should be presented to the Decision Maker respondent to explain and resolve it. If they are not able to do so, their responses to the core NTGR question with the inconsistency may be overridden by the findings from these supplemental probes. These situations should be handled on a case-by-case basis.

Early Applications of the Revised Nonresidential Self-Report Approach: Selected Findings

The nonresidential self-report approach summarized above was pre-tested in June and July of 2008. In addition to the completed pre-test cases, this methodology has subsequently been applied to a number of large projects involving multiple installations of a single measure for which the market is rapidly transforming. The circumstances surrounding these latter projects are very complex. General lessons learned from these early applications of the new methodology are presented below:

- While the core questions described above appear to support the calculation of internally consistent measures of free ridership using the NTG algorithm, the approach of **interviewing multiple project stakeholders and "triangulating" (i.e., comparing and combining) results** has proven to be effective in providing

the evidence needed to construct a consistent and plausible "story" behind each project, even the most complex ones. Interviews with participating vendors, account reps, and program reps have provided important context for the project. Information on general corporate energy and environmental practices, and/or past installations of similar equipment in regions without rebates have been used to establish standard practice, and allowed us to corroborate answers addressing measure installations that would be made in the program's absence.

- Valuable **cross-checking** of responses has occurred when facts discovered during interviews with project vendors and/or program staff have been presented as evidence during the interviews with primary decision makers in order to ensure their responses are consistent with previously-collected information. This process continues until a valid and plausible 'story' emerges.
- Another piece of the NTG puzzle is provided by the results of a detailed investigation of each company's **standard practice** in installing the subject technology. This investigation provides evidence of the customer's propensity to install the same equipment at the same time absent any program or rebate. Interviews are conducted with equipment vendors familiar with the company's past purchases of said technology without rebates, and with company officials involved in installing the same technology in other states that do not have rebate programs. Data regarding the company's historical installations of the energy efficient measure is also sought. In addition, copies of corporate energy and environmental policies are obtained.
- The **case study** procedure used for Standard – Very Large rigor projects involves constructing a chronology of the events and circumstances that led to the project's implementation, highlighting the program's importance vs. other factors. This chronology clearly lays out the time sequence of events surrounding the project, including the timing of program awareness and other key factors. Such information can provide key evidence of the program's influence, particularly if it comes at a critical time in the project before equipment purchase decisions are made.
- **Consistency checking procedures** have been very useful in uncovering conflicting answers that are then presented to the interviewee to for further explanation and resolution. These procedures are triggered in all cases, from the simplest projects to the most complex. Through the use of Computer-Aided Telephone Interview software, these procedures have now been automated, to ensure they are always invoked whenever inconsistent answers are given.
    - In one case, responses regarding other decision influences revealed during consistency checking made it clear that the need to adapt production in light of limited supplies of traditional raw materials played a much more significant role in the decision to implement than did program influences. Without this feedback loop, the inconsistency would probably not have been discovered, and the consequent importance assigned to the program influence would certainly have been overstated.

Conclusions

While the Nonresidential Self-Report methodology continues to evolve, the essential features of it are established and working well. Despite the simplistic label of "self report approach", this method incorporates evidence from multiple information sources in addition to the perspective of the key project decision maker. This evidence is used to provide context for the project, to cross-validate key results, and in some cases, as a direct input to the NTG scoring algorithm. By "triangulating" the results in this manner, a robust and defensible assessment of program influence (and associated free ridership) is reached.

References

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

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

Cronbach, L. J. (1982). Designing Evauation and Social Action Programs. San Francisco: Jossey-Bass.

DeVellis, R.F. (1991). *Scale Development: Theory and Applications*. Newbury Park, CA: Sage Publications, Inc.

Donaldson, Stewart I., Christina A. Christie, and Melvin Mark (Eds.). (2008). *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.

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.

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

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.

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). Bevery Hills, 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, 1972.

Weiss, Carol H. (1998). *Evaluation*. Upper Saddle River, New Jersey: Prentice Hall.
Guba, E. G. Toward a Methodology of Naturalistic Inquiry in Educational Evaluation (CSE Monographic Series in Evaluation No. 8). Los Angeles: Center for the Study of Evaluation.

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.