A Proposed New Framework for Evaluation in California: The Sampling Roadmap

Roger Wright and Tim Hennessy, RLW Analytics
Marian Brown, Southern California Edison
Nick Hall, TecMarket Works
Lori Megdal, Megdal & Associates
Ken Keating, Ken Keating & Associates

ABSTRACT

The California Evaluation Framework Study (2004) was undertaken to develop a new statewide framework for the evaluation of California energy efficiency programs. This paper focuses on the uncertainty and sampling segments of the project. The uncertainty chapter addresses three related topics: 1) statistical precision and bias, 2) integrating information from multiple evaluation studies, and 3) allocating evaluation resources among planned evaluation studies. The sampling chapter also has three main parts: 1) An overview of familiar techniques in simple random sampling, 2) a thorough discussion of stratified ratio estimation, and 3) a sampling roadmap that describes the chronological steps of a typical project. In addition, an example was developed illustrating the application of the roadmap to an actual study. This paper provides a concise overview of this material, highlighting some of its less familiar concepts and contrasting the new approach to the sampling approach given in the CADMAC protocols.

Overview

The California Evaluation Framework Study (2004) was undertaken to develop a new statewide framework for the evaluation of California energy efficiency programs. This paper focuses on the uncertainty and sampling chapters of the report. These chapters were designed to provide a self-contained discussion of techniques of statistical sampling that can be useful to a beginner in evaluation as well as to a more experienced study director or regulator. This paper provides a concise overview of this material.

The uncertainty chapter focuses on three related topics: 1) statistical precision and bias, 2) integrating information from multiple evaluation studies, and 3) allocating evaluation resources among planned evaluation studies.

The first section emphasizes that most commonly used techniques for reporting statistical precision – error bounds, confidence intervals, and relative precision – all assume that the results are free of bias. However, these methods can be very misleading if there is substantial bias in the findings. For example, it is misleading to report that the savings of a program have been measured within ±10% at the 90% level of confidence if there is a plausible risk that the results are biased by 25%.

Freedom from substantial bias is an underlying assumption throughout the remaining two sections of the uncertainty chapter. The second section shows how to combine the statistical precision of two or more evaluation studies. The third section discusses how to allocate

evaluation resources between two or more evaluation studies. A more flexible approach is recommended than the conventional requirement of ±10% relative precision at 90% confidence.

The sampling chapter also has three main sections: 1) An overview of familiar techniques in simple random sampling, 2) a rather thorough discussion of stratified ratio estimation, and 3) a sampling roadmap that describes the chronological steps of a typical project. In addition, an example illustrates the application of the sampling roadmap to an actual study.

The sampling chapter shows that the same equations can be used to plan a study using both simple random sampling and stratified ratio estimation. The expected statistical precision is determined by the a) population size, b) the sample size, and c) either the coefficient of variation of the target variable or the error ratio relating the target variable to the stratification variable. At the completion of the study, the coefficient of variation or error ratio can be calculated from the sample data to assist in planning subsequent studies.

Uncertainty

Bias and Statistical Precision

The discussion of bias builds on CADMAC (1998). Most conventional techniques of statistical inference, such as confidence intervals, rely on certain assumptions or models. As long as these assumptions or models are accurate, the confidence intervals are valid. But if the assumptions are materially wrong, then the confidence intervals can be very misleading. For example, with poor response rates or substantial measurement error, the results might have a bias of 25% or more, even though the traditional measure of statistical precision might be ±10% at the 90% level of confidence.

Evaluators face the following dilemma: the steps needed to reduce the risk of bias will usually increase the cost per sample point and therefore, if the project budget is fixed, force a smaller sample size. The smaller sample will generally yield a wider confidence interval than otherwise, but one that may be more valid.

Put differently, to the extent that the quality and soundness of a study is judged by the width of the confidence interval, there is a strong temptation to take shortcuts and put the saved resources into a larger sample.

The chapter goes on to discuss various sources of bias, including the following:

- Non-response and other forms of selection bias.
- Measurement bias.
- Erroneous specification of a statistical model.
- Choosing an inappropriate baseline.
- Self-selection of program participants.
- Misinterpretation of association as causal effects.

The chapter also discusses the difficulty of objectively quantifying the magnitude of the potential bias. Given the difficulties of an objective assessment, the principle investigators are generally in a better position than anyone else to assess the risk of bias. At the risk of sounding idealistic, we recommend that study directors attempt to help the reader (e.g., a policy maker or program planner) assess the danger of bias by pointing out any concrete evidence of a

breakdown of assumptions and by providing some assessment of the likely consequences. The example at the end of the sampling chapter illustrates how this might be done in a specific study.

Integrating the Results from Multiple Evaluation Studies

The second topic addressed in the uncertainty chapter is how to calculate the overall precision when several evaluation studies are involved. For example, a policy maker may want to: 1) estimate the total savings of a collection of programs, 2) compare and possibly combine the results of two independent studies that have assessed the savings of a given program, or 3) evaluate the statistical precision of the net savings of a program when one study has assessed the gross savings and a second, independent study has assessed the net-to-gross ratio.

Wright and Jacobson (1993) discussed a methodology for addressing questions such as these. In most applications of these techniques, the statistical precision of the combined result is better than the precision of the inputs. This is a consequence of a key assumption, namely that each study has provided unbiased results and valid confidence intervals and that the studies are statistically independent. Any bias in the studies can propagate throughout the analysis and seriously distort the conclusions.

This section also discusses the importance of converting a p-value to an error bound or relative precision. A p-value, sometimes labeled $Pr > |t|$, is commonly used in regression analysis to measure the statistical significance of a particular explanatory variable. Table 1 provides an example drawn from a billing analysis of a particular program. The primary parameter of interest is the regression coefficient associated with the indicator variable for program participation (Program). Since the associated p-value is very small (0.0078), an evaluator might conclude that the coefficient gives a highly reliable estimate of the savings.

Table 1. Converting a P-Value to an Error Bound or Relative Precision

Variable	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Error Bound*	*Relative Precision*
Intercept	1	493	163	3.03	0.0028	268	0.54
Pre-use	1	0.89	0.02	48.22	<.0001	0.03	0.03
Business	1	4,321	984	4.39	<.0001	1,618	0.37
Program	1	-536	199	-2.69	0.0078	328	0.61

In most evaluation applications, an error bound at the 90% level of confidence can usually be calculated as 1.645 times the reported standard error. The relative precision can be calculated as the error bound divided by the absolute value of the estimate. For example, in Table 1, a 90% confidence interval for the annual savings is 536 ± 328 units per participant. Moreover, the relative precision is only ±61%, which is rather poor. These results show that the p-value can give a misleading indication of the statistical precision of the results of a regression analysis.

Allocation of Resources to Evaluation

The uncertainty chapter also provides some guidance to the complex question of how much to spend on evaluating a portfolio of programs and how to allocate the spending to the individual programs in the portfolio. Many factors can influence the analysis, including:

1. The amount of savings expected from each program,
2. The uncertainty about the savings,
3. The unit cost of evaluating each sample project in the program,
4. The variability of savings in the population of projects in the program, as measured by the coefficient of variation or the error ratio,
5. Whether the program is expected to grow or shrink in the future,
6. How long it has been since the last evaluation and how much the program has changed in the interim.

The section leads off with a short discussion of Bayesian decision theory, power analysis, and propagation of uncertainty. Then it moves to two more conventional criteria: 1) fixed relative precision (e.g., requiring ±10% relative precision at the 90% level of confidence for each study), and 2) optimal allocation of the sample for assessing the overall savings of the portfolio. A methodology is provided that reflects the first four of the six factors listed above. The Framework report recommends that an evaluation planner should consider the remaining two factors and, if necessary, adjust the desired statistical precision or resulting sample sizes accordingly.

Sampling

The second of these two chapters discusses statistical sampling. In a typical evaluation study, data are collected and analysis is conducted for a sample of units, usually projects or customers, selected from a given population. By following statistical sampling methods, the data collection and analysis usually can be limited to a relatively small sample. For example, in an impact evaluation study, project-specific measurement and verification analysis might be carried out for a sample of 50 projects selected from the 1,200 projects implemented in the program in a given year.

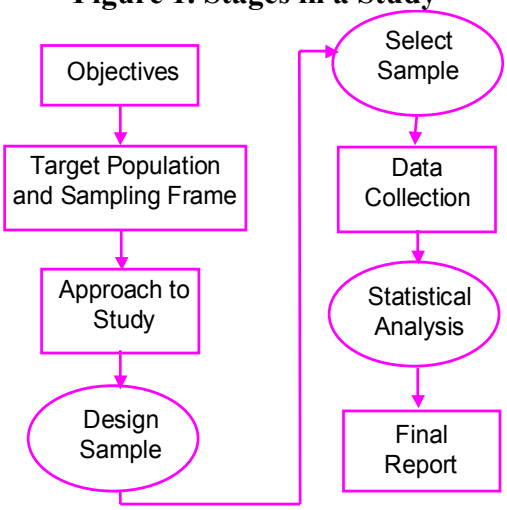
If the sample projects are selected following an efficient sample design, and if the data collection and site-specific analysis is free of substantial bias, then the statistical analysis can provide an essentially unbiased estimate of most population characteristics of interest and a good measure of the achieved statistical precision. The statistical analysis can also provide measures of population variability to guide future evaluation studies.

The effective use of sampling can help address the sources of bias discussed in the uncertainty chapter. By limiting data collection and analysis to a relatively small fraction of all projects, more attention can be devoted to each sample project. Some random measurement error is acceptable for each sample project, as long as the measurement error is small relative to the sampling variability. But systematic measurement bias should be minimized since it will propagate through the analysis. It is equally important to minimize other sources of bias such as non-response, self-selection, deliberate substitution, etc.

A study usually moves through the eight steps illustrated in Figure 1. The starting point is a clear understanding of the objectives of the study and the target population. For example, an impact evaluation study may be planned to measure the actual gross and net savings in the set of projects implemented in a given program in a given year. By contrast, in a baseline study the population may be the set of all nonresidential customers served by a given utility or the set of all vendors of a given appliance in a specified region.

The population database or sampling frame is a list of each unit in the population, with relevant information for each unit. If the target population is a set of projects implemented in a particular program year, then the population database is a list of each of these projects, together with supporting information for each project, developed from the program tracking system. The population database should provide the information needed to identify each unit in the population as well as a suitable measure of the size of each unit, e.g., the tracking estimate of the savings of each project in the program.

Figure 1. Stages in a Study



Simple Random Sampling

The statistical theory of simple random sampling is of interest because it is familiar and is the foundation of all other methods of sampling. Simple random sampling can be suitable in evaluation if the units in the population do not vary too much in size. This may be the case for some residential programs, and certain nonresidential programs if larger projects are suitably divided into subprojects that can be independently evaluated.

In simple random sampling, a sample of a given size, denoted n, is selected from the projects in the population following any randomized procedure in which all possible subsets of n projects are equally likely to be selected. A simple random sample can be selected by assigning a random number to each project listed in the sampling frame, sorting the sampling frame according to the random number, and designating the first n projects in the sorted list to be the sample.

Under simple random sampling, a single statistical equation is the basis for each of the following:

- Calculating the expected statistical precision for a specified sample size,
- Choosing the sample size to provide a desired statistical precision,
- Estimating the statistical precision from the sample data that have been collected, and
- Estimating the population parameters needed to plan future studies.

The probability is about 90% that the sample mean $\bar{y}$ *will fall within* $\pm 1.645\, sd(\bar{y})$ *of the population mean,* μ. Here $sd(\bar{y}) = \sqrt{1-\frac{n}{N}} \times \frac{\sigma}{\sqrt{n}}$ and σ is the standard deviation of the variable of interest (e.g., the true savings of each project) in the population.

Instead of working directly with σ, it is generally easier to work with the population coefficient of variation, $cv = \sigma/\mu$. If we define the expected relative precision at the 90% level of confidence to be $rp = 1.645\, sd(\bar{y})/\mu$, then we have $rp = 1.645\sqrt{1-\frac{n}{N}} \times \frac{cv}{\sqrt{n}}$.

The preceding equations can be used to guide the choice of the sample size under simple random sampling. The results have important policy implications. Assume that a fixed relative precision is required. If the population is large, then the sample size usually will be a small fraction of the population. But if the population is small, a substantial portion of the population may have to be sampled. This can have an important impact on the cost of evaluating small programs. In particular, the traditional criterion of requiring ±10% relative precision at the 90% level of confidence for each program imposes a disproportionately larger burden on small third-party programs than large utility programs.

For this reason, it sometimes may be desirable to consolidate small, similar programs for the purpose of evaluation. The target population could be taken to be the entire set of projects in a portfolio of related programs. A simple or stratified sample would be selected from the population. The annual savings of each sample project would be assessed using appropriate data collection and analysis techniques. Then the sample results would be used to estimate the average savings per project for all projects in the portfolio. With a sufficiently large number of projects within the portfolio, sampling will generally be cost effective and there should be adequate resources for proper attention to the risk of bias. Added information, e.g., a process evaluation, may be needed to assess the individual programs.

Stratified Ratio Estimation

Stratified ratio estimation combines a stratified sample design with a ratio estimator. Both stratification and ratio estimation take advantage of supporting information available for each project in the population. The chapter discuses an example based on an impact evaluation study of a particular C&I lighting retrofit program (Ledyard 2003). The program tracking system provides an estimate of the annual energy savings of each project, called the tracking kWh savings. Table 2 summarizes the tracking savings in the program population.

As the lower portion of Table 2 shows, the majority of the projects were relatively small but a few projects were very large. For example, over 1,000 of all 1,248 projects had tracking savings of 25,000 kWh or less. Moreover, the coefficient of variation of the tracking savings was quite large, 1.45. This indicated that the population coefficient of variation of the actual savings would also be rather large so simple random sampling would not be suitable, as will be shown below.

By contrast, stratified ratio estimation was very effective in this case. Stratifying by the tracking savings generally reduces the coefficient of variation of actual savings in each stratum thereby improving the statistical precision. Moreover, the sampling fraction can be varied from stratum to stratum to further improve the statistical precision. In particular, a relatively small

sample can be selected from the projects with small tracking savings, but the sample can be forced to include a higher proportion of the larger projects. In particular, the largest projects can be forced into the sample, i.e., included with certainty.

Table 2. Example – Summary of the kWh Savings in the Tracking System

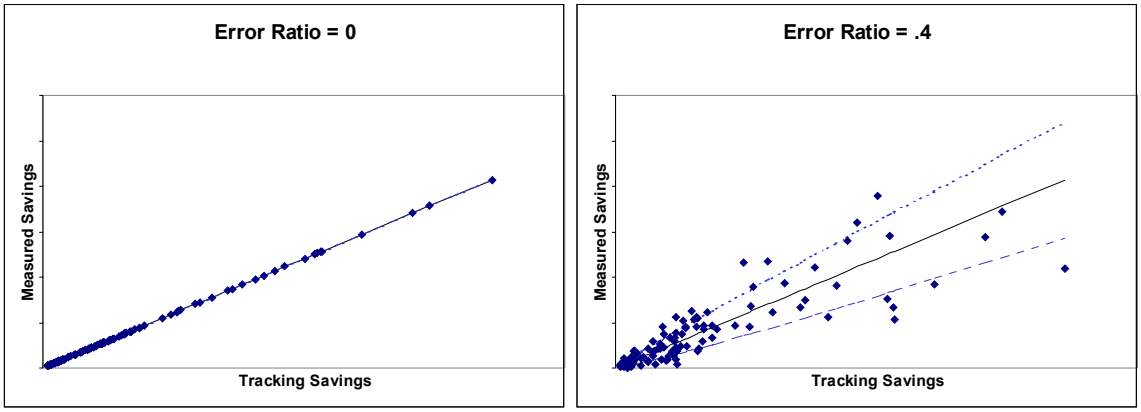
Population Size	1,248	Projects
Total Savings	20,119,315	kWh
Populaton Mean	16,121	kWh per Project
Standard Deviation	23,333	kWh per Project
Coefficient of Variation	1.45	
Minimum	48	kWh
Maximum	345,350	kWh

Max imum kWh	25,000	50,000	75,000	100,000	125,000	150,000	175,000	200,000	350,000
Number of Projects	1,011	152	49	15	16	1	2	1	1

The tracking estimates of savings can also be used in stratified ratio estimation. In impact evaluation, one ratio of interest is the realization rate, i.e., the ratio between the total gross annual savings of all projects in the population and the total tracking savings. The net to gross ratio is another ratio of interest. Our experience has been that ratio estimation can be used to estimate essentially all parameters of interest in evaluation.

To understand the potential advantage of stratified ratio estimation, suppose hypothetically that the actual savings of each project in the population is directly proportional to the savings recorded in the tracking system as illustrated in Example A of Figure 2. In this extreme example, the actual savings of each project is exactly 0.8 times the tracking estimate of savings. In other words, the tracking system systematically overstates the saving of each project by 20%. The realization rate, 0.8, is the slope of the line relating the actual savings to the tracking for every project. In this hypothetical case, the realization rate can be assessed perfectly by measuring the actual savings of any one project in the population.

Figure 2. Two Examples of the Association between Measured and Tracking Savings

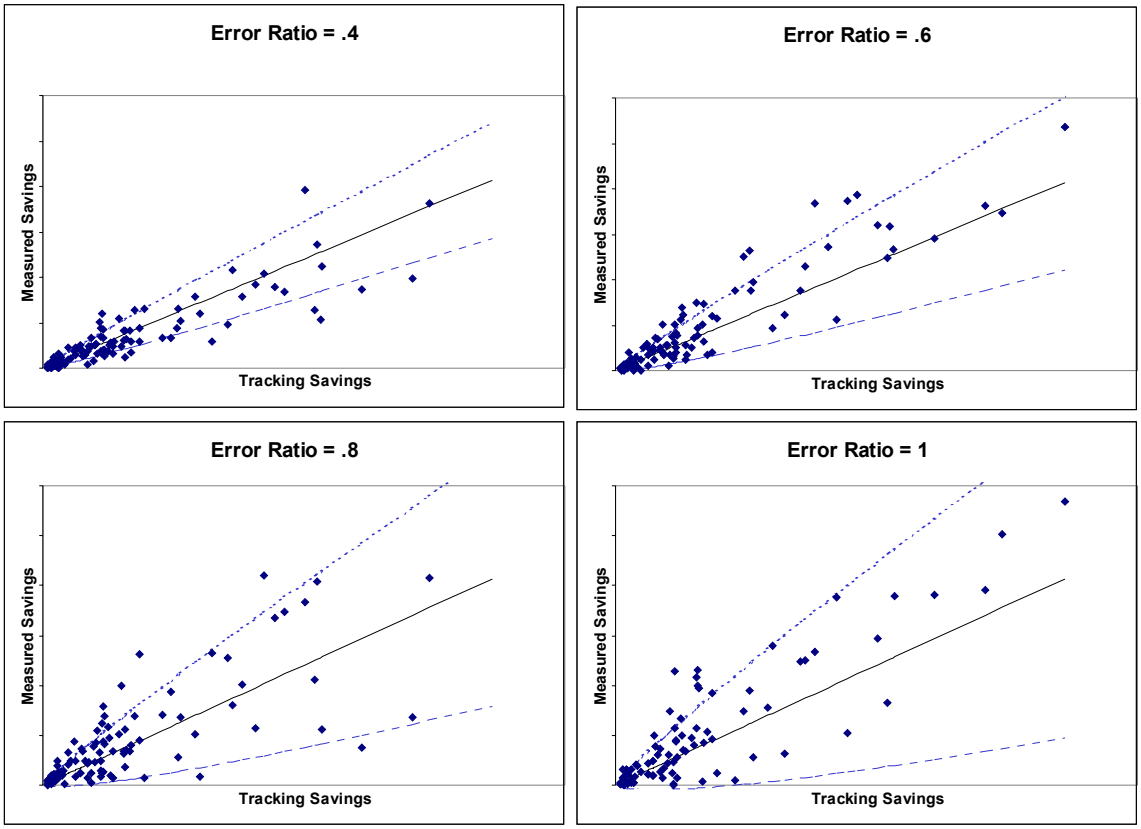


Example A **Example B**

In practice, of course, there is always some random error in the association between the actual and tracking savings. Example B of Figure 2 illustrates a more typical situation. In this case the tracking estimate of savings is a good but not perfect predictor of the actual savings of each project. Nevertheless, in this example the statistical precision can still be greatly improved by using stratified ratio estimation rather than simple random sampling.

For stratified ratio estimation, the key measure of the population variability is called the error ratio. The error ratio is a measure of the strength in the relationship between the numerator and denominator variables, e.g., the actual and tracking savings. Figure 3 shows several examples of error ratios ranging from 0.4 (a relatively strong relationship) to 1.0 (a weak relationship).

Figure 3. Examples of Different Error Ratios



The role of the error ratio (*er*) in stratified ratio estimation is virtually the same as the role of the coefficient of variation (*cv*) in simple random sampling. In particular, the expected relative precision at the 90% level of confidence can be calculated using the equation $rp = 1.645 \sqrt{1 - \frac{n}{N}} \times \frac{er}{\sqrt{n}}$.

In the example being discussed, a prior evaluation study of an earlier offering of this program indicated that the error ratio was about 0.4. Therefore, with stratified ratio estimation, a sample of 50 projects would give an expected relative precision of

$$rp = 1.645 \sqrt{1 - \frac{50}{1{,}248}} \times \frac{0.4}{\sqrt{50}} = 0.091.$$

This implies that under the stratified ratio estimation methodology, the expected relative precision would be about ±9% at the 90% level of confidence with a sample of 50 projects.

By contrast, using a coefficient of variation of 1.45, a simple random sample of 50 projects would have an expected relative precision of only ±33%. Put another way, in this example, simple random sampling would require a sample of 450 projects to yield the same expected precision as a stratified random sample of 50 projects.

The sampling chapter describes how to develop an efficiently stratified sample design. Table 3 shows a suitable sample design for the example. Stratum 1 consists of all projects with tracking savings of 10,128 kWh or less. Stratum 1 contains 694 projects with a total tracking savings of 2,892,887 kWh. The average tracking savings of these projects is 4,168 kWh per project. By contrast, stratum 5 consists of the largest 54 projects. These projects have a total tracking savings of 5,265,839 kWh and an average size of 97,516 kWh per project. The last column of Table 3 shows the desired number of sample projects in each stratum. In this type of sample design, the sample is usually allocated equally among the strata. In this example, a sample of 50 was planned with ten projects randomly selected from each of the five strata.

Table 3. An Efficiently Stratified Sample Design

	Tracking Savings				*Sample*
Stratum	***Projects***	***Max***	***Total***	***Average***	***Size***
1	694	10,128	2,892,887	4,168	10
2	253	18,981	3,502,474	13,844	10
3	151	35,341	3,979,634	26,355	10
4	96	62,056	4,478,481	46,651	10
5	54	345,350	5,265,839	97,516	10
Total	1,248		20,119,315	16,121	50

In this example, the recruiting and fieldwork were carried out mindful of the risks of bias that were discussed in the chapter on uncertainty. The fieldwork itself was designed to minimize measurement bias. The primary sample was carefully recruited to minimize the danger of non-response or self-selection bias. Designated backups were only used as a last resort. As shown in Table 4, the response rate was at least 70% in each stratum and almost 80% overall.

Table 4. The Disposition of the Sample

Stratum	Not Found	Refused	Scheduled	Total	Response Rate
1	1	2	10	13	77%
2	0	0	10	10	100%
3	1	3	10	14	71%
4	0	2	10	12	83%
5	2	2	10	14	71%
Total	4	9	50	63	79%

Table 5 shows the principal results for the gross realization rate. Based on the 50 sample sites, the gross realization rate was found to be 98.1%. In other words, the measured gross savings were estimated to be only 1.9% smaller than the tracking savings across all projects in

the population. The standard error was found to be 4.6%. This gave an error bound of ±7.5% at the 90% level of confidence. The 90% confidence interval for the gross realization rate in the population was from 90.6% to 105.7%. The relative precision was 7.7% at the 90% level of confidence. Finally, the error ratio estimated from the sample was found to be 29.8%, somewhat smaller than assumed in planning the study. The error ratio estimated in this study would be used to inform future sample designs for similar programs.

Table 5. Results for the Gross Realization Rate

Realization Rate	0.981
Standard Error	0.046
Error Bound	0.075
Low	0.906
High	1.057
Relative Precision	0.077
Estimated Error Ratio	0.298

The first step in developing these results was to calculate the case weights. The case weight (w) is simply the number of projects in the population in each stratum divided by the number of projects in the final sample in the corresponding stratum. The gross realization rate was calculated from the sample data using the equation

$$\hat{B} = \frac{\sum_{i=1}^{n} w_i \; y_i}{\sum_{i=1}^{n} w_i \; x_i} .$$

In this equation, the numerator is the sum of the products of the case weight and gross savings of the sample projects. The denominator is the sum of the products of the case weights and tracking savings. The standard error was calculated using the following equation with $e_i = y_i - \hat{B}\, x_i$:

$$se\left(\hat{B}\right) = \frac{\sqrt{\sum_{i=1}^{n} w_i \left(w_i - 1\right) e_i^2}}{\sum_{i=1}^{n} w_i \; x_i} .$$

Then the error bound at the 90% level of confidence was calculated using the equation $1.645\, se$, and the results were used to calculate the lower and upper boundaries of the 90% confidence interval. The relative precision was calculated by dividing the error bound by the estimated realization rate. Finally, the error ratio was calculated from the sample data using the following equation. Here, based on prior experience, we took $\gamma = 0.8$.

$$\hat{er} = \frac{\sqrt{\left(\sum_{i=1}^{n} w_i \; e_i^2 / x_i^{\gamma}\right)\left(\sum_{i=1}^{n} w_i \; x_i^{\gamma}\right)}}{\sum_{i=1}^{n} w_i \; y_i} .$$

The report discusses the risk of bias in this example. Ledyard (2003) addressed measurement bias. In the present study, the greatest risk was thought to be non-response bias arising from the replacement of some of the primary sample projects with backups, as summarized in Table 4. The first step was to estimate the amount of the tracking savings associated with the Not Found and Refused categories of program participants, based on the recruiting results. These estimates are shown in the first row of Table 6.

Table 6. Realization Rate by Recruiting Category

	Scheduled	Not Found	Refused	Total
kWh Tracking Savings	16,240,268	1,094,003	2,785,044	20,119,315
Gross Realization Rate	0.981	0.500	0.981	0.925
Est. Gross Savings	15,668,643	638,708	2,299,350	18,606,702

Then the realization rate was estimated in each of the three categories. In the Scheduled category, the realization rate from Table 5 was used, i.e., 0.981. In the Not Found category, it was assumed that these program participants might have gone out of business or were very seldom in their place of business. Consequently their operating hours might be shorter than recorded in the tracking system, leading to a lower realization rate than normal. From these considerations, it was felt that the realization rate might be lower than 0.98 but probably no smaller than 0.5 in this category.

In the Refused category, it was assumed that the realization rate was likely to be similar to that of the Scheduled category. Perhaps program participants refused to cooperate with the onsite audit because they were unhappy with their savings (low realization rate), or perhaps because they were very busy (high operating hours and therefore high realization rate). But it seemed most likely that their refusal was unrelated to their savings. Therefore, the realization rate in the Refused category was taken to be equal to the realization rate in the Scheduled category, i.e., 0.981.

In each of the three categories, the gross kWh savings was estimated by multiplying the estimated tracking savings by the estimated realization rate. Then the total gross kWh savings was calculated across all three categories, 18,606,702, and the overall realization rate was calculated as 0.925. By calculating the difference between this result and the realization rate previously calculated, shown in Table 5, it was concluded that the gross realization rate found from the sample of scheduled projects was likely to have a selection bias of about 0.056 or less.

It is important to recognize that the gross realization rates assumed for the Not Found and Refused categories were extremely subjective since very little information was available. In particular, there was no objective basis for assuming that the realization rate in the refused category was the same as in the Scheduled category. Sensitivity analysis was used to try different assumed realization rates for the Not Found and Refused categories. At worst, the realization rates could be assumed to be zero in these two categories. In this case the overall program realization rate was found to be only 0.779. In this worst possible scenario, the selection bias would have been 0.202, substantially greater than the error bound reported in Table 5.

In this example the response rate was almost 80%, which is rather high. Nevertheless, there was a notable risk of bias from non-response. This illustrates the importance of taking all available steps to minimize bias in implementing a study, and of providing a frank discussion about the dangers of bias in the final report.

Comparison to the CADMAC Protocols

The sampling chapters of the Framework report build on the CADMAC protocols that guided evaluation in California for many years. In particular, both CADMAC and the Framework report address the risk of bias. In Appendix J, CADMAC discussed a number of quality assurance issues that should be addressed in an evaluation study. These issues are closely related to the sources of potential bias that are discussed in the Framework report. Both CADMAC and the Framework report recommend that an evaluation report should provide a thorough discussion of the specific risks of bias that arise in a study.

However, the Framework report differs from the CADMAC protocols in several important ways:

1. CADMAC required that most studies be designed to meet ±10% relative precision at 90% confidence. The Framework report does not prescribe a set standard for statistical precision but recommends that both bias and statistical precision be considered in planning the study.
2. CADMAC did not specify which variable(s) were to be covered by the 10% / 90% standard but in practice the standard was often applied to the tracking estimate of savings. The Framework report provides a systematic approach for assessing the expected statistical precision of each important variable of interest.
3. CADMAC suggested that a census should be attempted when the program population has fewer than 300 units. The Framework report cautions about the risk of selection bias with an attempted census and recommends that sampling be used even for quite small populations.
4. CADMAC seems to suggest that each program be evaluated independently. The Framework report suggests that small, related programs may be combined in an evaluation study to reduce the evaluation burden of small programs and allow attention to mitigating bias..

References

The California Evaluation Framework (2004), California Public Utility Commission, http://www.calmac.org/calmac-filings.asp.

CADMAC (1998). *Quality Assurance Guidelines for Statistical, Engineering, and Self-Report Methods for Estimating DSM Program Impacts*. Appendix J.

Ledyard, Tom (2003). "Evaluating the Underserved Small C&I Market: Building a Bridge to Implementation." *International Energy Program Evaluation Conference*: Seattle, WA. pp. 627-637.

Wright, Roger and David Jacobson (1993). "A Methodology for Integration of Evaluation Studies." *National Demand-Side Management Conference*: Miami Beach, FL. pp 251-255.