Appendix J

Quality Assurance Guidelines For Statistical, Engineering, and Self-Report Methods for Estimating DSM Program Impacts

Revised June 1998

Table of Contents

Preface

The California Demand Side Management Advisory Committee (CADMAC) was established by the California Public Utilities Commission (CPUC). The CADMAC was charged with, among other things, the continuing overall development of the *Protocols and Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management Programs*, better known simply as the Measurement and Evaluation Protocols. Various subcommittees were also established at the same time that were responsible for addressing very specific methodological issues that arose in the implementation of the protocols. One of these subcommittees, the Retrofit Modeling Standards Subcommittee, was responsible for addressing issues surrounding statistical and engineering modeling and metering. In 1994, the original report, *Quality Assurance Guidelines for Statistical and Engineering Models*, was prepared at the request of this Subcommittee whose members at that time were:

John Peterson, Chair
Southern California Edison Company

Ben Bronfman
Consultant to the CPUC Division of Ratepayer Advocate

Sharim Chaudhury
Southern California Gas Company

Leon Clarke
Pacific Gas & Electric Company

Adrienne Kandel
California Energy Commission

Dean Schiffman
San Diego Gas & Electric Company

The authors of the original report were Richard Ridge, who at the time was employed by Pacific Consulting Services, Dan Violette, who at the time was employed by Xenergy, Inc., and Don Dohrman of ADM. In May of 1997, the QAG was revised to include a discussion of methods for estimating net-to-gross ratios based only on participant self reports. This addition, prepared by Katherine Randazzo and Richard Ridge, was needed since for some programs, e.g., industrial audits and rebates, the Protocols do not require a comparison group comprised of non-participants in order to estimate net impacts or net-to-gross ratios. The title of the revised version is *Quality Assurance Guidelines for Statistical, Engineering, and Self-Report Methods for Estimating DSM Program Impacts*.

This current revision has been prepared by Richard Ridge of Ridge & Associates, Katherine Randazzo of KVDR, Inc., Dave Baylon and Jonathan Heller of Ecotope, Inc., and Kevin Geraghty. This revision contains an entirely re-written Chapter 3, Quality Assurance Guidelines for Engineering Models, and important revisions to Chapter 4, Quality Assurance Guidelines for Estimating Net-To-Gross Ratios Using Participant Self Reports. These revision were done under the guidance of the Retrofit Modeling Standards Subcommittee whose current members are:

Pierre Landry, Chair
Southern California Edison Company

Randy Pozdena
Consultant to the CPUC Office of Ratepayer Advocate
EcoNorthwest

Jim Green
Southern California Gas Company

Chris Ann Dickerson
Pacific Gas & Electric Company

Adrienne Kandel
California Energy Commission

Dean Schiffman
San Diego Gas & Electric Company

1 Introduction

The California Public Utilities Commission (CPUC) recently adopted the *Protocols and Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management Programs (Protocols)* for the measurement and evaluation (M&E) of DSM programs. These guidelines focus on the critical elements of M&E such as load impact estimation models, sampling, and metering and are specific to various combinations of customer sectors, program types, and end uses. These standards are understood to be minimal and are in many cases quite general. For example, the Protocols state that the load impact models for commercial retrofit programs may be some variant of allowable CDA model types[1], or a calibrated engineering model, both possibly supplemented by an engineering simulation model. In addition, both participants and nonparticipants must be examined to estimate net program load impacts, and the sample sizes must be at least 350 for each group of non-residential customers or 200 for each group of residential customers. However, the Protocols are for the most part silent regarding such detailed methodological issues as the actual specification of CDA models, testing of statistical assumptions underlying CDA models, and power analysis. CE models and engineering models also lack any methodological guidance. Thus, simply adhering to these minimal standards contained in the Protocols is no guarantee that an analyst is doing a professionally respectable job.

While one could simply ask analysts to guarantee that they adhered to the methodological guidelines contained in standard textbooks, this may not be sufficiently reassuring either to utility or regulatory staff. Thus, rather than simply trust analysts to follow the guidance contained in the basic methodological textbooks, our preference has been to develop what is called the Quality Assurance Guidelines (QAG) that requires analysts to indicate specifically how they addressed basic methodological issues. This approach is clearly consistent with the white paper prepared by the ADSMP Subcommittee on Evaluation Standards and Guidelines and the *Program Evaluation Standards* prepared in 1994 by the Joint Committee on Standards for Educational Evaluation, in that it is not very prescriptive. That is, the Subcommittee members have thus far prepared *practice and reporting standards* rather than highly prescriptive *methodological standards*. Their preference has been to require analysts to describe how they addressed certain key issues rather than to require analysts to address these issues in a specific way. For example, while there are many varieties of regression-based analyses, there are very basic methodological issues that often come up, such as collinearity, and that must be addressed if one is to do a professionally respectable analysis. The guidelines only require analysts to test for collinearity but do not tell them how to test for it, and, if present, does not prescribe the appropriate remedy. This is the sort of guidance that

[1] For a more detailed definition of the various model types currently under discussion, please see "An Evaluation of Statistical and Engineering Models for Estimating Gross Energy Impacts" by Ridge et al., 1994.

occupies a position somewhere between the minimal standards represented by the Protocols and the highly detailed guidelines contained in basic methodological texts. The QAG also asks *where* certain information such as sample dispositions can be found in the report.

It follows that the QAG must focus on those methodological issues on which there is general agreement regarding their importance within the social science and engineering communities. The QAG will also refer analysts to texts in which more detailed guidance can be found regarding all the issues addressed. Adherence to such guidelines still allows the final models to be shaped by the interaction of the situation, the data and the analyst. It is this very interaction and the resulting plethora of legitimate methodological choices that prohibited the creation of a more detailed and prescriptive QAG.

The QAG can be used in several ways. First, they could be included as a part of every M&E request for proposals (RFP) so that prospective bidders will know that they will be held accountable for conducting a sound analysis. Second, utility project managers and regulators reviewing an evaluation report containing a completed QAG can quickly assess whether the analyst at least addressed the most basic methodological issues. This latter point is especially important since neither utilities nor regulators have the time or personnel to carefully scrutinize every written evaluation report let alone attempt to replicate the results of all these studies. Of course, the details of how they addressed these issues should be contained either in the very detailed documentation that would be contained in the technical appendix of any evaluation report or in the work papers. Finally, they can be used to create a common language to facilitate communication among utilities, regulators and consultants.

Analysts should not be expected to provide information on every model or approach they attempted during the analysis. One can get much of this detailed information from the analysis logs that every competent analyst should keep or from the computer output itself. Rather, the purpose of the QAG is to characterize what was typically done for the final models within each model type.

Included in the original report were the QAGs for statistical and engineering models. In May of 1997, the QAG was revised to include a discussion of methods for estimating net-to-gross ratios based only on participant self reports. This addition was needed since for some programs, e.g., industrial audits and rebates, the Protocols do not require a comparison group comprised of non participants in order to estimate net impacts or net-to-gross ratios.

There are several features of these QAGs that merit discussion. First, the issues addressed are issues that a variety of basic social science and engineering methodological texts also address. That is, there appears to be a consensus that these issues are important. Second, because the QAG is supposed to save time, it should not simply be an exact replication of what is in the report itself. On the other hand, for the same reason, it should not simply refer to the appropriate part(s) of the report. The answers, while brief, should provide enough information to reassure reviewers that a given methodological issue was recognized and dealt with in a professionally responsible

manner. Of course, only a pretest can determine whether this format will work. Finally, because some respondents may not be familiar some of the issues addressed or the terms used, references have been provided that should provide reasonably clear explanations.

2 Quality Assurance Guidelines for Statistical Models

2.1 Quality Assurance Guidelines for Conditional Demand Analysis (CDA) Models[2]

The QAG for CDA models is presented on the following pages. It is designed to cover the estimation of both net and gross impacts. With respect to net impacts, the issues addressed are well within the traditional research design framework involving the comparison of kWh consumption of participants and nonparticipants while attempting to control statistically for any compositional differences. Throughout the QAG, the observations participating in a regression model, or a sample, or in any other analysis framework are referred to simply as subjects, whether they are customers, accounts, or buildings and whether they are participant or comparison group members. Thus, how the questions are answered will depend on the type of study conducted.

This QAG should be completed for every CDA model type used in a given M&E study. However, a utility is not required to complete this form for every model attempted throughout the entire study within a given model type. One can get such detailed information from the analysis logs that every competent analyst should keep or from the computer output itself. Rather, in most cases, the purpose of the QAG is to characterize what was done for the final model(s) within a given model type. You should answer *each* of the questions briefly on separate pieces of paper.[3] Please keep your answers *brief*. You may refer the reviewer to specific sections of the evaluation report itself for more detail or perhaps for a complete and coherent answer to the question. Having said this, the reviewer should not have to piece together the answer from more than one section of the report. In other words, if the answer is in more than one section of the report, you must attempt to integrate the information from the report and provide the answer in a brief response. Remember, your summary should be much shorter than the discussion contained in the full report.

Note that some of these questions may not be relevant for a given study, thus making "not applicable (NA)" a legitimate response. For example, if you conducted a cross-sectional analysis, you should check "NA" for those questions relating to serial correlation.

---

[2] The definition of CDA is a collection of regression-based approaches that specify energy consumption as conditioned on any number of measured variables, but not a complete inventory of equipment or other demand sources. All of the regression-based approaches described in "An Evaluation of Statistical and Engineering Models for Estimating Gross Energy Impacts" by Ridge et al., 1994 qualify as CDA approaches. Other model types such as the statistical comparison method (SCM) and the calibrated engineering method (CEM) were considered sufficiently different from CDA models as to warrant their own guidelines.

[3] Each utility will provide a diskette containing all of the questions listed in the QAG. One can use this diskette to record all responses.

Finally, you may not be familiar with certain terms or concepts contained in the QAG. To assist you in completing the QAG, numbers are placed next to some of the section headings and/or questions that refer analysts to one or more methodological references in which the particular issue raised in the section or question is addressed. When appropriate, page numbers are provided. Of course, there are other references that could be used but the ones listed were considered adequate to describe the basic issues and their relevance as well as to provide methodological guidance in handling any related problems that may arise.

Quality Assurance Guidelines:

Conditional Demand Model Types

Date _____

Utility Program _____

Utility Project Manager_____

Lead Analyst _____

Employer _____

CPUC Study Identification Number _____

Sector(s) _____

Please indicate the sectors, programs, end uses, and measures for which estimates of gross and/or net impacts are provided. If impacts were estimated for other combinations of variables (e.g., weather zone or building type) please specify.

Are any of the impacts, adjusted for spillover? If yes, please describe.

_____

Period of Time Covered by the Analysis _____

Applicable Table(s) from M&E Protocols _____

Frequency of data (e.g., hourly, daily, monthly) _____

A. MODEL TYPES

1. Please check the model types used

a. Classic-Conditional Demand Analysis (C-CDA) using cross-sectional data and dummy variables to capture the impact of the program or installations _____

b. C-CDA using cross-sectional data and incorporating prior engineering estimates of impacts _____

c. C-CDA using cross-sectional time series (CSTS) data and dummy variables to capture the impact of the program or installations _____

d. C-CDA CSTS data and incorporating prior engineering estimates of impacts _____

e. Conditional Demand Analysis (CDA) using CSTS data and dummy variables to capture the impact of the program or installations _____

f. CDA using CSTS data and incorporating prior engineering estimates of impacts _____

g. CDA with pre/post design and dummy variables to capture the impact of the program or installations _____

h. CDA with pre/post design and incorporating prior engineering estimates of impacts _____

i. Other types of regression models used (please describe):

_____

_____

_____

B. MODELS

1. Please indicate where the forms of all the final models that were used can be found. Also, the forms of all the competing models that were used in the final stages of the analysis but were not selected as the final models are of interest. Please indicate where these can be found.

C. SAMPLE

1. Did you attempt to estimate models using the population of subjects or a sample? If a sample was used, describe the sample design. For example, what were stratification variables, if any, was the sample random, was the sample proportional, and how were the weights calculated?

1. What was the size of the outbound sample? For example, how many questionnaires were initially mailed out, telephone contacts attempted, on-sites attempted?

2. What was the size of the achieved sample? For example, how many completed questionnaires were returned, telephone interviews completed, on-sites completed?

3. What were the response rates for each of the major data collection efforts? For example, the response rate to a mail survey might be 50%, while for a telephone survey it might be 65%, and for on-site surveys it might be 85%.

4. Please indicate where more detail can be found on the sample dispositions for all major data collection efforts such as telephone interviews or mail surveys, on-site surveys, and billing data extractions. A sample disposition is simply a description of what happened to each effort to collect data (e.g., no telephone number, language barrier, refused, completed, etc., missing data in program tracking, billing or weather databases).

5. Describe any efforts to estimate the extent of non-response bias. For example, in order to measure any bias, did you compare the kWh consumption or other customer characteristics for respondents versus non respondents?

6. Describe your efforts to correct for non-response bias. For example, were respondents weighted in any way to correct for any bias?

7. Were procedures used to determine the size of the samples in order to achieve to specific levels of precision at given levels of confidence? If yes, what assumptions, i.e., expected variance or error ratio if model based sampling is used, or effect size in traditional power analysis, were used?

8. Describe *key* characteristics of subjects that you used in final models. For example, were they all installers of efficient equipment or were they simply exposed to some

treatment such as an audit? If residential were they single family or multi-family? What was the average income? What was the distribution across weather zones? If nonresidential, what building types and weather zones were represented?

D. DATA

1. Describe the data that were collected to support the analysis

2. Describe the source(s) and method(s) of collecting these data.

3. Indicate where the description can be found of how these data were manipulated in order to create the analysis datasets. Also, describe what screens were used to eliminate customers from the analysis and how many customers were eliminate as the result of each screen.

4. Where can all data collection instruments be found?

5. Where in the report can a flowchart be found illustrating the direct and indirect relationships of the data collected to each other and to the final estimates of impacts. If one is not available, please provide one.

E. SPECIFICATION AND ERROR

Misspecification

1. What were the initial specifications of the models and their rationale? If the specifications of these final models are different than these initial specifications, please explain what prompted the change. For example, were changes prompted by too much missing data for key variables, or the emergence of logical or theoretical inconsistencies?

2. Explain what you did to address the problem of misspecification. Describe the diagnostics carried out, the solutions attempted and their effects. If left untreated, please explain why.

Measurement Error

1. Were there substantial errors in measuring important independent variables? If so, what was done to minimize this problem. For example, was a weighted regression approach or an instrumental variables approach used?

Autocorrelation

1. If time series models were estimated, was autocorrelation a problem. If left uncorrected, biased estimates of standard errors may result. Under certain conditions, biased estimates of program impacts may also result. Please explain what you did to identify the problem in both the initial, intermediate, and final stages of the analysis

and what you did to mitigate its effects. Describe the diagnostics carried out, the solutions attempted and their effects. If left untreated, please explain why.

2. What was done to ensure the stability of the solution to serial correlation during the final estimation stages.

3. Were any checks done to determine if the pattern of autocorrelation differs by customer or building type and thus require a different type of treatment. For example, schools may have a different pattern than large office buildings. If differences were found among different sub groups, was autocorrelation treated differently for each of these groups?

4. Did the solution for autocorrelation negatively affect the solution for heteroskedasticity ? If so, what was done?

Heteroskedasticity

1. If heteroskedasticity was a problem, please explain what you did to mitigate its effects.

2. Describe the diagnostics carried out, the solutions attempted and their effects. If left untreated, please explain why.

3. Did the solution for heteroskedasticity negatively affect the solution for autocorrelation? If so, what was done?

4. If the solution to heteroskedasticity involved re-weighting of data, how did this weighting process interact with the relative weights or expansion weights developed on the basis of the sampling plan and nonresponse problems?

F. COLLINEARITY

1. Explain what you did to address the problem of collinearity as it may have surfaced in the initial, intermediate, and final stages of your analysis. Describe the diagnostics carried out, the solutions attempted and their effects.

2. What level of collinearity did you find acceptable and why? For example, some collinearity among regressors may be acceptable when the regressors are theoretically required, or when the regressors are necessary to represent a polytomy.

G. TESTS FOR EXOGENEITY

1. Tests to determine the exogeneity/endogeneity of variables are not routinely done, but, depending on the situation, they can be useful. For example, if any bias were suspected due to self-selection, such tests, described by Kennedy (1992) might be called for.

H. Influential Data

1. Describe any influential-data diagnostics that were performed in order to identify outliers?
2. If outliers were identified, how were they identified, how many were there, and how were they handled?

J. Missing Data

1. Describe how missing data were handled. For example, were cases with missing data dropped? Was mean substitution used to address the missing data problem or were other more acceptable techniques used?

K. Triangulation

1. If more than one estimate of impact is provided, how have the results been combined to form a single estimate?

L. Weather

1. Describe how weather normalization was handled. For example, were the kWh values weather-normalized prior to initiating the analysis or were models first estimated using the original kWh data and recorded temperature and later evaluated using long-run temperature data? In either case what was the source of the long-run weather?

2. Did the normalization adjust for heating degree-days only, cooling degree-days only, or both?

3. What degree-day base was used for heating and for cooling? If the base was customer-specific, how was the base selected?

4. Are there potential seasonal biases related to the pre- and post- period dates? For example, if one or more cooling seasons exist in the pre period while none exists in the post period, the savings estimates may be overestimated.

5. On a customer-specific basis, how was the choice made between a heating-only, cooling-only, or heating-cooling normalization model?

M. Engineering Priors

1. If prior engineering estimates of usage or savings were used in the models, what was the source(s) of the priors?

N. Precision

1. Where are the methods for calculation of key savings parameters and their standard error reported? For example, using standard statistical software, standard errors are always available on key parameters in a regression model while standard errors for

other parameters like net-to-gross ratios are often calculated during some post-processing of regression results.

O. Comparison Group

1. If a comparison group[4] *was not* used to help estimate gross savings, describe what was done to control for the effects of background variables such as economic and political activity that may account for any increase or decrease in consumption in addition to the DSM program itself.
2. If you used a comparison group to estimate either gross or net impacts, describe what was done to control for any compositional differences and any suspected self-selection bias.
3. If a comparison group was not used to estimate net impacts, please see Section 4 for a description of guidelines for estimating net-to-gross ratios using only the participant group's self-reports.

2.2 Quality Assurance Guidelines for Calibrated Engineering Methods (CEM)

2.2.1 Definition

Calibrated Engineering Methods use initial engineering estimates of impacts combined with a "statistical verification" step. This verification step produces an estimated realization rate. The application of these methods involves drawing a sample of program participants; then, an in-field metering or enhanced/in-depth engineering analysis based on other measures of customer consumption is conducted at each participant site. These analyses essentially "verify" or serve as "audited" values of the initial engineering estimates. A ratio is calculated between the audited values and initial engineering estimates. For example, if the audited values are, on average, 75% of the initial engineering estimates, then the ratio is .75. If the sample of customers is drawn randomly, then the best estimate of what the evaluator (or "auditor") would have found if the analysis could have been conducted on the entire population is .75 times the sum of the initial estimates for the population. One strength of this method is that, as long as the sampling is random, it is a relatively robust estimator. The types of assumptions required in the development of a regression model are not required by this method.[5]

---

[4] The M&E Protocols do not require a comparison group for estimating gross savings.

[5] It is important to note that CEM is different in concept from the calibration of engineering based energy simulation models. These engineering models provide estimates of levels of energy use and therefore the models are calibrated to observed data on usage levels (e.g., billing data or load research data). Impacts are then calculated by two runs of the engineering model—a baseline model run and a model run incorporating the energy efficiency measures.

A. PROGRAM RELATED AND MEASURE RELATED QUESTIONS

1. What energy efficiency measures are included in this program?

2. Were realization rates calculated for each measure? If no, what packages or combinations of measures were addressed?

3. What period of time is represented by the estimated impacts, i.e., what program interval is being estimated by this analysis?

B. SAMPLE AND SAMPLING (to be completed for each estimated realization rate)

1. What sample size was used to calculate the verified ratio?

2. Explain what procedures were used to help ensure that a random sample was drawn.

3. Were any tests or comparisons made to examine whether the drawn sample was "representative" of the population of participants? If yes, please explain.

4. Were any adjustments to the sampling plan made to make the sample more "representative?" If yes, please explain.

5. Was a stratified sampling procedure used? If yes, please describe briefly, with rationale for stratification choices.

6. Were procedures used to size the samples to target specific precision levels at a given level of confidence? If yes, what assumptions, i.e., expected variance or error ratio in the case of model based sampling, were used?

C. VERIFICATION/MEASUREMENT METHOD

1. What procedures were used to verify the kWh and kW impacts for specific sites?
   a) pre/post end-use interval metering? If yes, what was the duration?
   b) spot watt metering pre and interval metering post? If yes, what was the duration?
   c) pre/post spot watt metering with post run-time metering? If yes, what was the duration?
   d) engineering analyses? If yes, please explain.

2. Were weather related/seasonal effects estimated? If yes, please explain how.

3. Were interaction effects addressed, for example the heating penalty associated with lighting efficiency improvements? If yes, please explain.

4. Were changes made to the baseline energy use from which the impacts are estimated, e.g., were changes made to account for burned out lights, expansions of space (adding on a new wing), changes in use of space, etc.? If yes, please explain.

5. While on-site, were issues of snap back, free-ridership, spillover addressed with the customer? If yes, please explain.

D. REPRESENTATION OF RESULTS

1. Within the sample, was one realization rate estimated for the entire sample, or were realization rates allowed to vary by any factor (e.g., magnitude of savings, type of building, total energy consumption)? If yes, please explain.

2. After the sample was drawn, were the strata boundaries and associated case weights adjusted to reflect the most current information on the population of participants? If yes, please explain.

3. Was an analysis of influential data points conducted? If yes, please explain.
   [Note to reviewers: For example, the sample sizes are usually small and it is easy to exclude each observation and re-estimate the realization rates to determine whether a single observation greatly influences the realization rate estimate. Similarly, potential outliers, e.g., savings estimates more that three standard deviations above the mean, can be excluded and the realization rate re-estimated. The distribution of savings tends to be a skewed distribution often with a few sites having extremely large savings estimates. A realization rate is a straight line fitting process, the y/x ratio as the slope, and the effect of several points well outside the general range of observations could influence the estimate. Many interesting estimation issues are involved, e.g., do the sets of observations essentially come from different distributions; that is, are they generated from a different underlying process and therefore belong in a different analysis.]

4. If influential data points were identified, were they analyzed to see if they were unique cases, i.e., did not fall within the definition of the program being analyzed? If yes, please explain.

5. Were any drawn sample sites dropped from the analysis for any reason? If yes, please explain.

6. Did your analysis present the mean, median, standard deviation and an example precision level and confidence interval for each realization rate estimate?

2.3 Quality Assurance Guidelines for STATISTICAL COMPARISON METHODS (SCM)

2.3.1 Definition

These approaches, sometimes termed “simple” comparison approaches, involve pre/ post comparisons of energy use among program participants. This involves subtracting the annual consumption in the “post” period from the annual consumption in the “pre” period. However, before this subtraction is done, the energy consumption data must first be weather-normalized: the effects of atypical weather are removed to produce what is called normalized annual energy consumption (NAC). The simple equation for calculating gross impacts for participants is presented below.

$$\text{Savings} = \text{NAC}_{\text{Pre}} - \text{NAC}_{\text{Post}}$$

One of the more commonly used methods for weather normalization is PRISM, which, like many of these comparison methods, typically does not use any data other than consumption data and weather data.

A. PROGRAM EFFICIENCY MEASURES

1. What energy efficiency measures are included in this program?
2. What proportion of program participants had each energy efficiency measure?
3. Were savings estimated separately for different participant groups? If yes, how were the different groups defined (by measures, by timing of participation, geographically, by characteristics known from the customer information system, etc.)?
4. What was the timing of program participation for the estimated savings?

B. COMPARISON GROUP

With respect to the estimates of gross savings, the M&E Protocols do not require a comparison group. The rationale for its exclusion is provided on pages 2-4 to 2-6 of *An Evaluation of Statistical and Engineering Models for Estimating Gross Energy Savings*. Thus, the SCM as defined does not include a comparison group. However, some have recommended the use of a comparison group to help control for exogenous changes related to prices, political factors, technological changes, or any systematic bias in the weather normalization procedure. Of course, a comparison group is often required for estimating net savings. The questions below are relevant if one used a comparison group for estimating either gross or net savings.

1. Was a comparison group used in the analysis?
2. How was the comparison group defined?

3. How were pre- and post- periods defined for comparison group customers?
4. Were any tests or comparisons made of the similarity between the participants and comparison group in the "pre" period? (yes/no) If yes, describe.
5. Were any adjustments made for differences between participants and the comparison group? If yes, describe.

C. SAMPLE AND SAMPLING

1. How many participating (and how many comparison) customers were used to estimate the savings?
2. Were these customers selected from the total pool of participating (nonparticipating) customers:
    - randomly
    - by census
    - by other means (explain)?
3. What screens were used to eliminate customers from the analysis?
4. How many participants (comparison cases) were eliminated as a result of each screen?
5. Were any tests or comparisons made to examine whether the drawn sample was "representative" of the population of participants (comparison population)? (explain)
6. Was a stratified sample used? (yes/no) If yes, how were strata defined and how was the allocation to strata determined?
7. Was the sample weighted in the analysis? If yes, what was the basis for the weighting?

D. WEATHER NORMALIZATION

1. What weather normalization model was used?
2. What time period defined the "normal" weather?
3. What was the source of the weather data used for the analysis?
4. What pre- and post-participation dates were included in the analysis?
5. Are there potential seasonal biases related to the pre- and post- period dates?

6. Did the normalization adjust for heating degree-days only, cooling degree-days only, or both?
7. On a customer-specific basis, how was the choice made between a heating-only, cooling-only, or heating-cooling normalization model?
8. What degree-day base was used for heating and for cooling? If the base was customer-specific, how was the base selected?
9. What accuracy measures are reported for the normalization model fits?

E. DIAGNOSTICS AND ACCURACY

1. Were the normalized savings examined for outliers?
2. How were cases identified as outliers handled?
3. Were any comparisons or tests made of the sensitivity of the results to inclusion or exclusion of outliers? If yes, describe.
4. Is the standard error of the estimated savings reported?
5. Is a confidence interval for the estimate reported? If yes, at what confidence level?
6. Is there a discussion of potential biases in the analysis?

3 Quality Assurance Guidelines for Engineering Models[6]

3.1 INTRODUCTION

These quality assurance guidelines under the California Protocols are intended to establish a consistent basis for conducting engineering reviews of utility programs aimed at providing energy conservation in various sectors. It is the purpose of this set of guidelines to develop and expand on the conditions under which engineering analysis including sampling guidelines, would be deemed to provide sufficient demonstration of the energy savings resulting from technologies installed under utility energy efficiency programs.

Currently, the Protocols only allow for the use of simplified engineering models (e.g. algorithms) or more complex building simulation models (e.g. DOE2) for specific programs and end uses as the primary method for estimating *gross* impacts. Table 1 illustrates the Protocol tables, the DSM program, the end uses, and the allowable engineering techniques.

While an engineering analysis which estimates gross impacts can be conducted on virtually any program for any utility, the Protocols focus on the more complex programs in which statistical billing analyses perform poorly; e.g. in cases where the number of participants is relatively low, the size of the sector is small, the expected savings are small, or the diversity of participation and/or applied measures is such that no consistent representation can be made. For the most part, these problems arise in the analysis of industrial and commercial programs.

In these cases, the review should be conducted on a case-by-case basis in which the engineering analysis applied to particular participants is reviewed or re-evaluated, and savings estimates are justified with respect to the observed billing records or to a detailed engineering review (based on the characteristics of the particular participant and applied measure). In some programs, this would indicate a more rigorous sampling design, while in other programs additional resources to allow participants to be reviewed individually are required.

---

[6] While the focus in this section is on the two energy simulation models most commonly used in California, DOE2 and Micropas, the issues addressed and the questions asked are also generally applicable to other models. Until a later study is conducted that addresses any possible issues that are unique to models other than DOE2 and Micropas, analysts should rely on the guidelines for engineering models contained in this report.

Table 1: Engineering Technique Allowed in Protocols by Table, Program, and End Use.

Table	Program	End Use	Engineering Technique
C-3A	Residential Appliance Efficiency Incentives Program	Lighting	Simplified Engineering Model
C-4	Commercial Energy Efficiency Incentives Program	HVAC	Building Simulation Model
C-5	Industrial Energy Efficiency Incentives Program	Indoor Lighting; Motors; Industrial Process	Simplified Engineering Model
C-6	Agricultural Energy Efficiency Incentives Program	Pumping	Simplified Engineering Model
C-7	Residential New Construction	Whole Building	Building Simulation Model
C-8	Nonresidential New Construction	Whole Building	Building Simulation Model
C-9	Miscellaneous Efficiency Program Measures	Miscellaneous	Simplified Engineering Model
C-12	Nonresidential Fuel Substitution	Water Heating; HVAC; Process; Water Pumping	Simplified Engineering Model; Building Simulation Model

3.2 ENGINEERING REVIEW AND ANALYSIS

Although engineering reviews are commonly used for many purposes, this set of guidelines will focus on the establishment of the relationship between claimed savings and verified or actual savings. This relationship is inherent in the program design and evaluation protocols. Thus, a sampling methodology which generalizes the relationship between utility calculations for claimed savings and actual savings could be developed that would minimize errors introduced by differing calculation methodologies used by individual utilities.

Because of the relatively intensive analysis required for any one participant or any one program, the use of engineering analysis as an evaluation tool should be carefully allocated to those programs or problems that can be modeled with relatively limited

samples. Where access or participant identification is problematic, engineering analysis often is either difficult or impossible to apply.

An engineering review is intended to establish the determinants of consumption in the context of the installed equipment at the participant level. The engineering algorithms used to determine the equipment sizes and schedules in the design process are applied to the evaluation process. Usually, some adaptation is required; however, standard software tools are available for the more complex engineering problems.

The engineering analysis seeks to determine the following:

- What are the determinants of consumption in this particular process or building?
- How can we understand (and quantify) these determinants?
- How does the equipment installed by the conservation program impact total consumption by changing these determinants?

For the evaluation of energy savings of conservation programs, this analysis focuses on the change in energy use from a well-defined base case. The engineering analysis is used to compare the predicted energy savings from utility and manufacturing representatives, installers, or other parties with the observed conditions at the site or with detailed information collected by account representatives or other utility personnel as part of the program implementation. This process ferrets out the errors in original savings calculations, as well as the bias and approximations that must be made during the planning process.

3.2.1 Data Collection

The data collection methodology must be designed carefully to characterize all of the components needed to determine consumption related to a particular end use. The principal goal is to describe the level of savings achieved by the measures installed versus either the pre-installation conditions or some standardized set of components that would have been installed in the absence of the utility program. In either case, an engineering review would focus on the determinants of consumption for that particular end use and the resulting impact of the installed versus the base case. For comparison purposes, the same algorithm would be used in both cases and the output from the analysis would be the change in modeled consumption between the two conditions.

The information collected should focus on the measures or actions taken due to the utility program *and* the base case information needed for the savings calculation. This should include occupancy information, changes in occupancy conditions, and a detailed engineering description of the conditions effected by the utility program. In the industrial sector, for example, a detailed audit and review of the processes of each building is usually essential. This should include:

1. The output of the particular production line;

2. The overall productivity of the plant;
3. Changes due to efficiency improvements.

For most building-based conservation programs, information about building components, number of occupants and hours of occupancy, and mechanical equipment must be established. This is particularly important in developing energy simulations. Other equipment that interacts with the HVAC system (such as lighting, building service water, office machinery, and other components) must also be described in some detail. The result of combining all of this information is a detailed audit, which is aimed at establishing the determinants of energy consumption and the necessary inputs to produce a valid engineering analysis.

Utility files can often be the only source of base case information. Unfortunately, many utility files are either incomplete or inadequate. In general, careful review of these files is important to establish the information needs of the engineering review.

There is a tendency, during data collection and engineering analysis, to use proprietary software which is difficult to review in its general form or as it relates to a particular building or engineering problem. While there are times when this is acceptable (such as when evaluating particular components in the industrial sectors) it is generally a poor idea to base an evaluation or a savings claim on software which is inherently difficult to review. It is not very helpful when the computer code is published, but summary information explaining the techniques used is important to establish the credibility of such techniques.

The remainder of this section focuses on different aspects of, and techniques used in, engineering analysis of these conservation programs.

3.2.2 Base Case

Regardless of the evaluation technique, one of the most important tasks for the evaluators of each project site is the establishment and documentation of the base case. Depending on the measure and program, the base case is either the situation as it existed prior to the installation of the energy conservation measure, or some estimation of what would have been installed in the absence of the conservation program. Load impacts are derived from the difference between the energy use following installation of the utility-sponsored measures and the energy use under the base case.

In most retrofit programs, the connected load, efficiency, and hours of operation of the existing equipment which is to be replaced or upgraded are required to determine the base case. This establishes the amount of energy needed to accomplish a certain task originally, which is then compared to the energy needed to accomplish the same task after the measures are installed.

There are two basic retrofit conditions. The first involves situations in which there are *no* efficiency standards for the installed equipment. This requires that the initial utility estimate of savings made during the implementation of the program be based on information gathered and documented *prior* to the installation of the conservation measures, such as nameplate ratings and hours of operation of the equipment which is being replaced. In more complex situations, one-time or short-term metering may have been conducted by the utility. It is not usually sufficient to rely on the memory of operations staff for these initial estimates as this may be biased by expectations of greatly improved performance. Without documentation of the prior condition, it is often not possible to justify energy savings calculations. This requirement does not apply to those DSM programs for which there are so-called *deemed* baseline values that have been agreed to by the utility and the CPUC. The second involves situations in which there *are* efficiency standards for the installed equipment. In such situations, the usage associated with the efficiency standard may serve as an acceptable base case.

Of course, in either of the two conditions described above, one may wish to investigate partial freeridership. In such cases, an alternative base case could be established using information provided by the operations staff.

In the case of new construction, a comparison of the installed measure must be made to an estimated base case comprised of what would have been installed in the absence of the program. The definition of the base case can be based on plans drawn up before the involvement of the program or the average efficiency of equipment recently installed outside the program at the participating site. Another approach can base the base case on an industry recognized "standard practice" for that particular end use. A definition for "standard practice" for a particular end use can be developed from code regulations or the average efficiency of equipment being installed in similar buildings in other locations.

Any such base case is clearly difficult to establish since it is theoretical in nature. In these cases, the burden is on the utility to develop a base case that can be justified. Without a believable base case which is less efficient than the installed measures, it is not possible to claim that any energy savings have been achieved.

In the industrial sector, where equipment is tied to productivity or output, the impact of a modified production process could result in an increase in output. The base case for the savings calculation is more complex in these cases and is discussed in detail in Section 4.

As with the other energy use calculations, the base case should be calculated in as straightforward a manner as possible and in a manner that parallels the methods used for the incented or as-built cases.

If there are no significant interactions with other energy uses, then a simple engineering calculation should be sufficient to establish the base case. If interactions are present, then a simulation will be required. The same simulation must be used for the base case and the new condition, containing all of the same assumptions (unless it can be demonstrated that the utility program had a direct effect on those assumptions).

3.2.3 Simple non-interactive measures

Many conservation measures involve changes in the connected load or hours of operation of an end use, which has little or no interaction with other energy uses in the building. In these cases, a simple engineering model can be employed to calculate the impact on total energy demand and total consumption related to the changes made. There is no reason to use a complex simulation model, since the energy impacts can be more precisely described by simpler engineering algorithms. This analysis method is most often seen in lighting measures, where there is an absolute watt reduction as a result of changing equipment. In such cases, review of nameplate capacities can be sufficient to establish the impact of the measure. There are other situations, however, when before- and after-retrofit metering must be applied to determine the absolute load and the change in consumption.

In the most straightforward situation, the principal impact of the conservation measure is to reduce the connected load associated with a particular end use. For example, when exterior lighting measures are applied to fixtures lacking lighting control or scheduling, the energy saved is calculated using simply the difference between the initial wattage and the final wattage after the installation of the measure. This figure is multiplied by a presumed or measured duty cycle, based either on *a priori* hours of operation or metered or monitored schedules. In this example, no changes in energy use in any other building component or operation could be expected. Therefore, a simple fixture audit would provide the information required to calculate total savings. Similarly, in agricultural pumping, the connected load impact of a single-speed motor used in a pumping application can account for virtually all energy savings. However, given the wide variation in duty cycles over the course of a given season, assumptions or data must be collected to demonstrate savings based either on actual observed duty cycles in the program year or some normalized duty cycle averaged over several crop years.

In general the primary source for the base case and other determinants of consumption are files generated by the utility documenting the characteristics and initial conditions of the participant. If these files are properly constructed, this simplified engineering analysis can be limited to verification of the assumptions and engineering contained in the files. It is difficult to imagine that this would ever be simply a file review. In virtually all cases a field review of some sort would be required to verify the engineering basis of the savings estimates. The resulting file would then be a combination of the evaluator's review of the particular customer and the program summary of the utility's calculations used in developing the savings estimates for the individual customer.

3.2.4 Simulations of Load Interactions

The use of engineering simulations has been a major component of engineering analysis in California programs. This is a situation that should be approached with some caution. Many conservation measures will have an impact on more than one energy use

component in the building (for example, where an interaction between the energy use of cooling equipment and the reduced internal heat resulting from lighting conservation in a commercial building). A typical methodology for evaluation is to generate a simulation of this interaction so that the net savings associated with particular interactive measures can be established.

When developing such simulations, it is not only important to obtain accurate information about individual pieces of equipment, but also to obtain a complete view of that particular system's design and operation. The inputs from occupancy characteristics (such as hours of occupancy, number of occupants, and process loads) can have a larger effect on the total HVAC use than the components that are affected by the energy conservation measure.

Simulation strategies usually solve this through the use of a calibration to overall energy bills. This calibration can be quite complicated, particularly if the initial data collected about occupancy and other issues is incomplete or inaccurate. Often, this calibration step requires more effort than the initial set up of the simulation. Furthermore, simulations often have a relatively limited usefulness. If the energy conservation measure is aimed at a domestic hot water system, for example, the utility of a detailed simulation of the entire building is negligible. For typical buildings, this type of simulation is most useful to analyze interactions between HVAC systems and other building loads.

3.2.4.1 Non-Residential Buildings

For non-residential building simulations, DOE2 is considered a standard. It provides a fairly straightforward engineering simulation that attempts to simulate climate, building equipment, and building shell, and the interactions among these components. It is not the only energy simulation tool available, nor is it always the most applicable to particular conditions. The DOE2 program is well developed around a fairly approximate building load in algorithm with a very detailed equipment simulation package, which can account for many of the operating characteristics of heating, cooling and ventilating equipment. The simulation requires many hundreds of inputs which must be generated from audits or engineering reviews. Assumptions or default parameters are used for variables that cannot be developed in this way. The particular combination of these assumptions will have a significant impact on the absolute size of the energy prediction from the simulation. The default assumptions can have a greater impact on energy use predictions than the audit results and should be documented as carefully as the engineering inputs derived from the analysis.

The DOE2 program is part of an entire class of simulations (e.g., TRACE, TRANSYS, HAP, etc.), each with its own strengths and weaknesses for particular analysis applications. DOE2, for example, has a fairly cursory treatment of the building shell and allows relatively little direct evaluation of complex building shell improvements such as may be found in the residential sector. DOE2 typically uses approximations for interactions between various heat loss components such as ground connections and has a relatively approximate treatment of solar interactions with the building load.

For most non-residential buildings, this simplification is quite acceptable. By far the most significant energy use factor in these buildings will be the result of internal activities and/or equipment efficiencies and configurations in the building itself. Carefully and correctly describing these factors is the most significant factor in understanding the energy use of the building. While the DOE2 program has limitations in this area, it is designed directly around this problem and is among the most likely simulation tools to establish the relationships between building loads, climate and equipment.

3.2.4.2 Residential Buildings

In the residential sector, where energy use is largely determined by climate conditions and building shell characteristics, the DOE2 simulation is often too approximate. The importance of equipment efficiency in residences is often secondary to occupant behavior (such as thermostat setpoint), local microclimate variation, and building heat loss rate or solar heat gain through windows and other surfaces. Depending on the nature of the question, simulations aimed at the residential sector can often be simplified substantially. There are numerous simulations in California which are meant to solve this problem (such as CALPAS, CALRES, and MicroPAS) as well as simulations distributed nationally which are similarly constructed around the development of a careful understanding of building loads caused by climate. Generally, one of these alternatives should be preferred to DOE2 for use in the residential sector or when the measure is significantly impacted by climate or envelope considerations.

3.2.4.3 Bin Method Analysis

There are also numerous simplified simulation programs based on the bin method that can be used for any construction type. The bin method actually summarizes the climate of a particular place based on the probability of particular temperatures, and analyzes the building based on its performance within normal temperature ranges for a particular site. This is often a very convenient method, since most heating and cooling equipment is rated, and manufacturing specifications are evaluated using this approximation method. Part load curves are typically expressed as equipment behavior within temperature bin widths of about 5°*F*. Thus, a program that establishes consumption within these temperature bins can be easily adapted to equipment part load curves of this sort.

The advantage of these programs is that they require substantially fewer assumptions; however bin simulations are also less accurate and should not be used if a significant amount of detailed information is available. Bin programs are designed around simulating building load and the interactions between building load and building equipment. Occupancy schedules, lighting schedules, lighting loads, process loads, domestic hot water loads, and other loads not directly related to the climate or equipment performance are inputs into the simulation. These can be measured directly, via the connected load evaluation, or can be assumed from general observation of particular building types. Where the analysis is focused on equipment performance with relatively

little occupant-introduced complexity (e.g., warehouses, small retail buildings, etc.), the bin method might be acceptable.

3.2.5 Model Calibration

Quite often, individual simulation programs (especially proprietary programs that develop inputs from audits or short term data collection) are based on a long list of assumptions. These assumptions are often based on long term experience or on detailed metering performed on unrelated buildings. It is important to realize that no matter how much detailed data is collected, a complex simulation such as DOE2 will require several thousand assumptions, and even the more simple models require several hundred. The characteristics associated with these assumptions can often be as important in determining the results of the simulation as the detailed review of the equipment or building shell itself.

It is for this reason that the calibration of model outputs to the observed bills for a particular building can be very misleading. The simulator utilizes numerous assumptions which can have major impacts on building energy use, but which cannot be directly known from audits or other direct short-term observations. Such things as ventilation rates, incidental infiltration rates, thermostat setpoints by individual occupants, etc. can all be adjusted to achieve a match to an arbitrary billing standard. It may well be that these are not the assumptions that lead to the correct building specification. Other errors may also be found elsewhere in the simulation's description of equipment configuration, building shell configuration, or interactions between distribution systems and equipment.

The standard reference material on the proper simulation of buildings and building processes is quite limited. Experienced modelers are important; however, documenting the general simulation procedure in detail is also important so that reviewers can assess the combination of assumptions that have been used. Once the crucial assumptions necessary to describe the measure and its interaction with the building or process have been derived, the details of the exact match to the billing calibration are secondary. Usually demonstrating that such a match can be made is all that is required. Even then, the function of this step is largely to establish the veracity of the simulator, not the simulation.

It is therefore necessary that the assumptions used in the simulation are clearly accessible to the evaluator. Some complex simulation programs are such that a huge number of simulations are done inside a complicated batch file arrangement. These can be arranged in such a way that the individual critical assumptions effecting energy use are effectively hidden from the reviewers. It can be quite cumbersome to open individual input files and search for the assumptions that were made. The evaluator should provide hard copies of all input parameters used to model each structure (e.g., a print of the BDL files for simulations using DOE-2). Finally, there should not be a large amount of adjustments made to standard assumptions for individual building files to match them to bills without explicit documentation of those adjustments.

The calibration step involves the evaluation of bills. As with any billing analysis (Section 3), the analyst should be careful to calibrate the model with billing results that are as free of error and complete as possible. Because of variation in weather and meter cycles, it is very desirable to calibrate simulations with bills covering an entire year. Shorter periods can be used if there are no alternatives but careful comparisons using explicit weather and occupancy data are necessary to assert that the calibration is meaningful.

3.3 Billing Analysis

While billing or regression analysis techniques are often used as an alternative to a detailed engineering approach, they can also be important in supplementing or calibrating the engineering evaluation. These tools can be effective when used with care to inform a more detailed engineering analysis.

3.3.1 Conditional Demand Analysis or Statistically-Adjusted Engineering Estimation

Conditional Demand Analysis (CDA), or Statistically-Adjusted Engineering (SAE) Estimation are statistical approaches that involve the use of multiple regression with binary "dummy" variables or variables providing engineering estimates of energy use. The model typically employs energy use as the dependent variable and a variety of other variables hypothesized to affect energy use (including the installation of the conservation measure) as the independent variables.

Such an approach asserts that energy savings can be due to a variety of factors, including more efficient equipment or changes in consumer behaviors. If a building is unchanged, but the occupant sets the cooling setpoint five degrees higher than during the previous cooling season, energy savings are realized. As a corollary, the assumption is that consumer use patterns can sometimes negate the benefits of increased efficiency in building technologies. To understand the amount of energy conservation achieved, one need only extract a large enough representative sample of the population and test to see whether less energy is used by this group after controlling for the usual exogenous variables (such as other changes in the building that may affect energy use, climate, economic activity, etc.).

The problem with CDA, using binary dummy variables, is that it often requires a very large data set. To be statistically valid, the model requires at least as many degrees of freedom as there are variables in the regression. This is true even in the case of regression models in which a preliminary regression is used to estimate an intermediate variable. This variable is then used as an input to the final regression. When this is done, there must be at least as many degrees of freedom in the data set as <u>all</u> of the variables used in both the preliminary and final regression models.

3.3.2 Load Shape Analysis

This technique uses individual bills in combination with a detailed engineering description of a specific building to evaluate overall energy conservation effects. This differs from typical regression analysis in that the actual billing analysis is conducted on a case-by-case basis, using the bills and other information about the building to disaggregate consumption and provide estimates of particular components of the energy bill. A typical example of this technique is the PRISM® degree day program used in the residential sector to disaggregate utility bills into their major components using a regression technique.

These programs can be quite useful; however, they have severe limitations concerning the need for good billing data and relatively consistent energy use over a period of approximately one year. This works well in the residential sector and can be useful in the industrial sector in cases where plant output and operation remain unchanged. However, in the non-residential sector, billing analysis is useful to calibrate other engineering analyses in cases where simulations or other detailed analysis can be informed by total energy use estimates. In some situations where detailed energy usage information is available for loads, the need to disaggregate end uses is minimal (such as a review of gas bills where only space heating is present).

The load shape analysis can proceed with a regression or other statistical approach to understand the relationship between climate variables or audit variables and monthly energy use from bills. Quite often, the billing analysis is the only real basis for establishing whether complex engineering approaches are adequate for the particular building or process. In this event, the billing analysis is just one step, which must be supplemented by additional engineering analysis to explain the findings or to establish the veracity of the engineering analysis compared to actual consumption.

3.3.3 Data Collection

For any billing analysis methodology, the data collection process is straightforward. However, there are several pitfalls, which should be highlighted:

1. The relationship between the collected bills and the process or building to be reviewed must be understood. It is not unusual for there to be substantial errors in both assigning bills to particular customers or end uses, and in readings or billing estimates that are part of the billing record. For the utility, bills are an accounting record used to invoice individual customers. The accuracy of particular meter readings is not vital to this use, as long as the errors can be corrected in a subsequent billing. Thus, if a particular bill is estimated because of some internal billing process, it can be easily corrected upon the next actual reading. For a multivariate billing analysis used to develop climate parameters, however, this can be extremely difficult to interpret. The analyst must carefully survey the billing record to ensure that this relationship is as error-free as possible. This is a tedious but straightforward process, since quite often these bills show back-to-

back deviations, with one being substantially higher than expected followed by one substantially lower than expected (or vice-versa). Correcting for these involves the use of either a "smoothing" algorithm (if sufficient data is available) or graphical processes in which the judgment of the analyst is relied upon to smooth the bills over a period of erratic records.

2. Missing bills are another common problem. Often, the size of the billing record is insufficient at the beginning of a building's operation. Depending on the nature of the analysis used and measures to be reviewed, this could be a serious problem that would result in the individual case being excluded from further analysis. The period for which billing records are available might be sufficient to establish the level of consumption. It is important to realize that either of these outcomes can be found, but if a billing analysis depends on annual or seasonal variation, this variation should be observable in the billing record if it is to be used as part of a billing analysis. For example, space heating and cooling loads in the residential sector demand a fairly complete one year billing record in order to track seasonal variations and estimate heating and/or cooling effects. It is often difficult or impossible to develop calibrations for particular end uses from billing records for this reason.

3. Important end uses (such as space conditioning or domestic hot water in residential construction, or dominant process loads in industrial applications) can use disaggregations of the bills as the basis for calibrating particular engineering models. For the most part, however, the bills are much coarser than is required for this purpose. They can be indicative of the accuracy of other methods but are not likely to be useful without additional information about the billing components or process. The record of production in an industrial process could explain some or all of the variations in the billing analysis and help normalize the billing record to a relevant production level in an industrial plant.

4. In calibrating the engineering analysis using billing analysis, it is particularly helpful to remember that occupancy changes are likely to contaminate the analysis and need to be accounted for directly. Since the sampling is based on an efficient sample of a savings or consumption variable, it is unlikely that errors due to occupancy changes in the residential or commercial sector can be assumed to be random and unbiased. Thus, a direct accounting of these changes should be attempted. In some sectors, especially where the meter is attached to a single end use (such as an agricultural pumping station), the bills can be used as a direct measure of savings as long as the efforts to normalize consumption across the base case and program year periods are made.

3.4 Deferred Savings (Production Increments)

A special situation exists for calculating an accurate base case when the utility program involves production equipment that may exist primarily in industrial facilities but also in commercial facilities such restaurants and stores. In the industrial and commercial

sectors, energy conservation programs often target production processes such as plastic extrusion and cooking. In the most straightforward cases, the program affects only the efficiency of production, so the energy savings are simply the amount of energy which was originally required to produce the product minus the energy required to produce the same amount of product at the new efficiency level. However, often the project will affect not only the efficiency of production, but the rate of production as well, allowing the same plant to produce more product than before the installation. In the most extreme cases, the installation of new production equipment enables the customer to increase production so much that it actually uses more energy than before the installation of the energy conservation measure (this can be true even if the efficiency of production per unit of product is significantly improved). At the same time, there are external market fluctuations which may cause a plant to produce more product in a given year, causing an increase in production after the installation of new equipment which was not related to the conservation program at all.

It is not the intent of the protocols to support increase in production or facility modernization for its own sake. A clear increase in production efficiency for the present increment of production must be demonstrated. If the production is increased or new production capacity is added then the standards for inclusion in the shared savings incentives programs must increase. There are two general classes of these "production increments" or "deferred savings" that must be considered:

3.4.1 New production line or new facility

Some measures are implemented in entirely new customer facilities or as part of an addition to a customer's facility, e.g., a new building containing entirely new manufacturing equipment. In this case, there is no pre-installation output level. The evaluation must use the observed output level in the post-installation period. Note that the procedures described above for determining the appropriate baseline for energy and demand as well as calculating the kWh and kW impacts also apply to the new-facility circumstance.

While the baseline requirements for the new facility parallel the NRNC protocols, new industrial or commercial production should be filed under the Energy Efficiency Incentives for the industrial (IEEI) or commercial (CEEI) sectors. In these cases however the utility or the customer should submit a case from which savings are calculated. This should include vendors, technologies and performance information so that an engineering evaluation of the base case can be undertaken in reviewing the new facility claims.

3.4.2 Improvements in existing production lines, proving the rebates did not cause the increase in production

The default assumption for the remaining rebate measures will be that they caused the change in post-installation output. Two forms of evidence will constitute sufficient proof

that this assumption is wrong. These two forms of evidence are as follows:

1. **Customer Testimonial in Application File**. The utility may place in the customer's file a letter, on customer letterhead, dated prior to the date of installation or the date of the application for the incentive (whichever is earlier), which states that the customer had planned to change the plant's output. Many customers consider their future output levels to be highly confidential, so it is not necessary that the output level be quantified. If a letter is present in the file, it takes precedence over any other data gathered in the post-period evaluation.

2. **Decision-Maker Interview in Post-Installation Period**. If no letter is present, a second form of proof may be sought during the evaluation. An interview may be conducted with a member of the customer's staff who is responsible for planning the output level of the measure-affected systems. The interview will consist of a battery of questions aimed at estimating the probability that output would have reached the same level in the post-period in the absence of the rebate measure. A high probability will constitute proof that the measure did not cause the change in output level.

If a utility chooses not to collect either of these two forms of evidence, the utility must assume that the rebate measure caused the increase in the post-installation output level.

3.4.3 Adjusting for Changes in Output in Gross Savings Estimates

Engineering estimates of savings require estimates of energy use for measure-affected systems under two conditions. The first is the baseline condition, which may be represented by the pre-installation performance characteristics of the affected equipment or in the case of new facilities by an assessment of a current practice standard. In cases where the increased production is the result of the efficiency improvement then at a minimum the new increment of production must be compared to a baseline that represents the current practice for new production in the industry. Thus, a separate estimate is also required for the post-installation conditions. Using these two pieces of information, savings are calculated as baseline consumption minus post-installation consumption. One major variable in this calculation is the output level of the affected system.

The treatment of output level in the calculation must reflect the determination of whether the measure caused the post-installation change in output level. There are two possible cases.

1. If the measure **caused the change** in output, gross savings are defined to be:

   (Consumption of the affected systems in the post-installation conditions, assuming that systems were operated to achieve the pre-installation output level) minus (consumption that would have occurred if the unimproved system had been used to achieve the pre-installation output level).

2. If the measure **did not cause the change**, gross savings are defined to be:

(Consumption of the affected systems in the post-installation conditions at the observed post-installation output level) minus (consumption that would have occurred if the unimproved system had been used to achieve the post-installation output level).

3.5 SAMPLING

As they are currently written, the Protocols emphasize the need for fairly large and extensive samples to overcome the difficulties introduced by diverse building populations and use patterns. The guidelines presented, while they pre-judge the results of random sampling, suggest that the authors of the Protocols are interested in a simple random sample that is large enough to smooth out the variances within the sample itself.

Given an optimum stratification and sample design, the size of the final sample required could be reduced dramatically. However, if the principal goal of the evaluation is to fit a regression model with a large number of parameters, then the sample size still needs to be quite large. The marginal cost of those sample sizes is not great, since the evaluation depends on the collection, cleaning and use of utility bills with relatively cursory secondary information which is easily gathered from telephone surveys or similar data-gathering efforts.

If an engineering evaluation is to be conducted, then the marginal cost of gathering information on individual customers becomes much higher. In this event, the need for well-designed samples, which are stratified to accommodate both the range of consumption and the end use activities, is essential. Careful engineering analysis demands much greater resources per site than simple billing analysis. The use of focused and well-designed sampling plans makes it feasible to use engineering analysis as the basis for evaluating even very large programs.

A more detailed discussion of sampling methods is presented in Attachment 1.

3.5.1 Stratified sampling

A powerful, yet fairly simple, method of using auxiliary information is to construct a stratified sample. A good stratified sampling plan can reduce the sampling requirements for a given program by a factor of ten when compared to a simple random sampling. That is, if you chose your sites by simple random sampling, like beans from a jar, you would need to evaluate four to five hundred sites for the same level of precision that forty or fifty sites chosen through a stratified sampling plan would deliver.

Careful development of a stratified sample can avoid its inherent pitfalls. In general, it is best if the stratification criteria are derived directly from the measurement most desired: claimed kWh savings, for example. Stratifications that are unrelated to the target variable can actually decrease the precision of estimates relative to simple random

sampling. This implies that if you wish to verify claimed kWh savings, you should avoid stratifying on geographical region or equipment type and stratify solely on claimed savings.

Of course, it is tempting to stratify on equipment type as well as on claimed savings, so as to derive verification ratios or realization rates for different equipment types; but it must be appreciated how expensive this practice can be in terms of the precision of the overall estimate. An unfocused stratification plan can fail in its principal objective - verifying program claimed savings. Verification ratios for specific equipment classes can still be estimated given that these equipment classes do not enter into stratification criteria; but they are in the nature of 'bonus' information; the overriding objective is not compromised to aid in their estimation.

A different sort of pitfall is created by sampling plans that arbitrarily exclude entire classes of sites from the sample. Typical subjects of such exclusion would be sites which are small in some sense or hard or inconvenient to evaluate. The problem with such *a priori* exclusions is that the resulting program estimates are generally biased. It is only common sense that the collective claimed savings accounted for by the smallest consumers can not be verified if those customers are excluded from the sample. The intuition behind such arbitrary exclusions is, of course, that the verifiers wish to concentrate their efforts on more "important" sites. However, a properly constructed sampling plan typically samples only a few of the smallest sites, implying, in effect, that each sampled "little" site is standing in for a great many other unsampled little sites and is hence as important in the overall estimation as "big" sites sampled at much higher rates. The low sampling rates for little sites imply that a sampling plan free of arbitrary exclusions (and resulting biases) entails little or no extra effort.

The Dalenius-Hodges formulae, for example, constitute a method for determining optimal stratification boundaries and sample sizes within each stratum. Optimal in this context means finding the way to achieve a needed precision level with the fewest sample points, or, if the cost of assessing different kinds of sites differs, at the lowest weighted cost.

An unexpected pitfall with optimal stratification schemes is pushing them too hard. A finely calibrated sampling plan with many savings-based strata and minimal sample sizes within each stratum is more vulnerable to problems such as data contamination and non-response than a simpler sampling plan with fewer strata. The optimizing formulae work on the assumption that the true variability of savings within each stratum is known; in fact we assume that variability of claimed savings is a good proxy; if it is not, our sophisticated sampling plan can perform badly. In practice, a maximum of six strata is probably enough for program evaluation, even if the formulae suggest that we could reduce needed sample sizes further with ten or twelve strata.

3.5.2 Model-assisted sampling

Classic stratified sampling uses available auxiliary information in the sample design phase, to determine stratification boundaries. Model-assisted sampling, a later development, uses auxiliary information both in designing a sample and in calculation of subsequent estimates. The fuller use of auxiliary information means that model-assisted sampling can potentially outperform classic stratified sampling. The cost of these potential benefits is the making of more assumptions and greater complexity of estimation. If our assumptions turn out to have been far from the truth, the resultant estimates, while still statistically valid, can have low precision. The assumptions that set model-assisted sampling apart from classic stratified sampling are that there exists a regression relationship between the variable of interest (verified per-site savings, say) and known auxiliary variables (claimed per site savings) which explains a high proportion of the variability across sites in the target variable.

As with classic Dalenius-Hodges stratified sampling, auxiliary information is also used in the sample design phase. An optimal sample plan sets the likelihood of selecting a given site proportional to the standard deviation of its regression error. In practice, it is quite difficult to come up with a plan for drawing a sample of fixed size without replacement, which satisfies the above criterion. One response to this quandary is to accept a random sample size; another response is to reintroduce stratified sampling.

A straightforward near-optimal stratified sample can be designed, in which selection probability is nearly proportional to regression error standard deviation. The stratification variable is not the magnitude of claimed savings at a given site, as in the classic stratified sample case discussed above, but rather the assumed relative standard deviation of the regression error at that site. It turns out also that the preferred scheme for determining stratum boundaries does not follow the classic Dalenius-Hodges criterion, although, as with Dalenius-Hodges it is optimal once these boundaries are set to sample equal, or nearly equal numbers of sites from each constructed stratum. As with classic stratified sampling, with the above scheme it is costly to use unrelated variables as additional stratification criteria.

The principal use of a model-assisted methodology is to provide a method for expanding information gathered on a small sample to a larger sample that has been designed to meet the sampling criteria. The advantage of this type of sampling is that it provides a basis for subdividing the sample for purposes of performing complex and expensive analysis on a sub-set of the sample. This has usually been either sub-metering of particular buildings for purposes of gathering inputs to an engineering analysis such as light schedules or cooling loads. This technique has been applied as a “double ratio analysis” in some evaluations, but has been largely misapplied (if widely accepted) in engineering analysis conducted in the non-residential sectors.

The second area where this technique has been employed (with considerably less effect) is in the use of elaborate simulation calibration exercises that single out a few buildings for detailed simulation with feedback from utility billings. This process usually allows

the modeling contractor to demonstrate prowess with the simulation tool but little else is accomplished. Relatively rarely is this double sample applied correctly so that the variance calculations (thus the significance tests and confidence intervals) are modified to recognize that the stage-wise model input has the effect of changing (increasing) the variance of the estimate.

3.6 Guidelines Summary

The following elements are required to accurately measure and defend savings estimates for installed conservation measures.

1. Data necessary to establish the base case energy use of existing equipment that is to be replaced or modified must be clearly documented *before* the installation of the new equipment. In some cases, it may be possible to reconstruct the base case from company files, but the memory of operations staff cannot be relied upon for schedules or energy use. As a result, the utility should insure that the base case is carefully documented for any program that is to be evaluated with an engineering review.

2. For purposes of calculating load impacts, simulation of base case must use the same simulation program and the same assumptions that are used to simulate the energy use of the new condition. Assumptions may only be altered for the new conditions if it can be clearly demonstrated that the program had a direct effect on that aspect of the energy use.

3. The burden is on the utility to prove that any increase in production of an industrial process site was due to factors unrelated to the program if savings are to be calculated on production increments. Furthermore, they must establish a new equipment base case for that particular industry if new production is added as a part of the project.

4. Simple engineering algorithms can only be used to calculate load impacts when there are no important load interactions between the installed measures and other energy-using processes in the building.

5. Simulations should be used when there are significant interactions between the conservation measure implemented and other energy uses in the building.

6. The simulation tool should be well matched to the building type or end use. DOE2 should only be used for commercial buildings with complex interactions of space conditioning equipment with other loads. Other simulation programs are acceptable, but the analyst is required to document the applicability of the simulation to the sector and analysis conducted. Bin methods or simpler hourly simulations should be used for the residential sector or for other cases dominated by climate and building shell interactions.

7. Standard assumptions affecting energy use that are used in the simulations must be clearly stated and not buried in individual building files. Any variations from these standard assumptions for individual buildings must be documented and justified.

8. Automated input and setup routines used in the development of load estimates for particular applications must be documented, including the default parameter settings used by the analyst or assumed by the standardized set-up routines.

9. An industrial process application must show that an increase in production between the base case and the improved case is traceable to market conditions and not to production improvements brought on by the incentive measures. If this case is not made, then load impacts shall be calculated using the production prior to the installation of the measures.

10. The production efficiency base case shall be developed from current practice in the industry, not from the previous conditions in the production facility.

11. Stratified sampling is encouraged in conducting engineering analysis. Samples should meet the confidence intervals and significance requirements of the protocols. To achieve this, somewhat more rigorous criteria should be employed in the sample design. Generally, a six level stratification would be considered the maximum number of strata for a simple stratified random sample. Model assisted samples may employ more strata (particularly in two-way sampling).

ATTACHMENT 1:

Sampling Guidelines and Procedures

Introduction

Conceptually, it is useful to think of two separate components to any sampling exercise: the sample design and the plan of subsequent analysis adopted. The sample design and the analysis plan are interdependent, because the choice of analytical approach will imply a preferred sample design, but independent in that a chosen analytical approach can be applied to a wide variety of sample designs. Sample design is defined as a plan for picking a sample from a population. More abstractly, a sample design implies the assignment of a selection probability to every subset of a population.

Consider a simple example -- the drawing of 10 separate sites at random from a program population of 100. This sample design assigns a selection probability of zero to every subset of the 100 which does not contain 10 elements; and assigns an equal (and very small) selection probability to every 10-member subset of the 100. A sample design implies a selection probability for every individual element (site) in the population; but such individual selection probabilities do not, in general, determine a sample design since selection events for different sites need not be independent. The fact that one site has been selected may (or may not) affect the probability of another site being selected. It certainly does so in the case of the simple random sample of 10 from a population of 100. The unconditional probability of individual site selection is 1/10; the probability of a site selection given that one has already been chosen is 9/99, and so forth; the probability of selection for the remaining sites declines as more and more sites are chosen.

A contrasting sample design in which site selection probabilities are independent is the list-sequential Bernoulli scheme. This method involves assigning a 10% selection probability to each site, one by one. For example, draw a number at random between 0 and 1; and set the selection criteria to sites with an assigned number that is less than or equal to .1. The selection probability for each site is 10%, as with the simple random sample design, but selection probabilities are now independent, and the total number in the sample is now a random variable with a binomial distribution.

Good sample designs and statistical analysis plans for program evaluation have some things in common: they make full use of already available information to, in effect, form prior guesses about evaluated savings at particular sites; and they arrange to sample more heavily amongst sites about which there is more uncertainty. The uncertainty obviously has to do with what we already know about the sites we are sampling. In the case of program evaluation, we typically know a fair amount -- at least the class and size of equipment installed at a given site and the magnitude of claimed savings (in kWh, kW, or therms). The precise meaning of "uncertainty" used depends on the plan of subsequent analysis, but is usually expressed as a standard deviation. Here we lay out two orthodox, fairly simple and much-tested sample design and statistical analysis approaches. We also suggest some variations and more complex approaches, which might be useful.

Classical Stratified Sampling and Mean Estimation

The sampling approach with the longest pedigree is stratified sampling, coupled with direct estimation of total program savings as a weighted sum of individual stratum means. Suppose,

for the sake of illustration, that we are interested in evaluating claimed kWh savings for a program. We would proceed by sorting the sites in the program database by magnitude of kWh claimed savings, and breaking up this sorted list into some predetermined number of strata. An individual stratum would be defined as those sites with claimed kWh savings lying in some interval.

For the moment, leave aside the question of how many strata we would choose and where we would draw the boundaries between adjacent strata. Given that our population of program sites has been divided into strata, and each site has been assigned to a stratum, how do we proceed to draw our sample? The principle that we sample most heavily where we are least certain is the key. Consider the standard deviation of evaluated savings within each population stratum around mean evaluated savings in that stratum. This is a measure of the variability within the stratum of evaluated savings. If we knew, or could guess at this number for all strata, then it would be optimal (in the sense of having the smallest sampling variance) to sample in each stratum proportionately to that stratum's standard deviation. The higher the standard deviation, the more sample points, relatively, we take from that stratum.

This scheme for determining sample sizes in each stratum, given a fixed overall sample size, is called the Neyman allocation, after its originator. The Neyman allocation formula can also be used to determine the minimum number of sample points needed to attain a desired variance of final estimates. In practice, of course, we do not know the variance of evaluated savings across sites within each stratum; but if we are willing to assume that the variance of claimed savings within each stratum is roughly equal to, or proportional to, variance of evaluated savings, then we can plug these claimed savings variances into the Neyman Allocation formula to get an optimal sample design, given the strata.

In general, one observes that savings standard deviations within program strata grow at least proportionately to the average size of claimed savings within those strata. Thus the Neyman allocation directs us to sample the strata containing large-savings sites with much higher relative frequencies than the strata containing small claimed-savings sites. Given how skewed program population claimed savings are, in fact, it is common to "census" the largest stratum, that is, to select every site into the sample.

Strata Design

So far we have not discussed the question of how to create the strata, but taken them as given. For a fixed number of strata, there is an optimal way of creating stratum boundaries if we know the distribution that governs evaluated savings in the population (and if we plan to use the Neyman allocation to determine our sample once we have defined the strata). The (approximately optimal) rule is to cumulate equal amounts of the square root of the density function ($\sum_{h=1}^{H} N_h \bar{y}_{s_h}$) into each stratum. This is the Dalenius-Hodges formula.

Of course, we do not actually know the distribution of evaluated savings within the program population, any more than we know the within-stratum variances of evaluated savings used by the Neyman allocation. If we did know this, there would be no need to do any evaluating. The solution, as previously, is to use the distribution of claimed savings within the program population. In practice it can be tricky to estimate a density function from a finite and highly

skewed population. A procedure often used is to assume some parametric family of skewed distributions (such as the exponential distribution) and fit that to the program population of claimed savings before invoking the Dalenius-Hodges method. The Dalenius-Hodges method, coupled with the Neyman allocation, implies that equal, or nearly equal, numbers of sites should be selected from each stratum. This is a useful check to see that one's routines are working properly.

Once the sample is drawn and evaluations are conducted, we can apply standard formulas to estimate population-level savings, or equivalently, per-site savings, and an accompanying variance. The formula used to estimate population total evaluated savings is $\sum_{h=1}^{H} N_h \bar{y}_{s_h}$. In this formula, there are presumed to be *H* strata; N_h is the program population count in the *h*-th stratum; and $\bar{y}_{s_h}$ is the sample mean of evaluated savings for the *h*-th stratum. A mathematically equivalent and more general form of the same formula is $\sum_{k=1}^{n} y_k / \pi_k$. In this formula, *n* is the total sample size, y_k is evaluated savings at the *k*-th site; and π_k is the probability of selection for the *k*-th site. Each sampled site is weighted in the overall estimate by the inverse of its prior likelihood of selection. This formula can be turned into the preceding, more specialized form by noting that the selection probability π_k for sampled sites in the *h*-th stratum is n_h / N_h, the number sampled from the *h*-th stratum divided by the population count from the *h*-th stratum.

Ratio Estimators

One way to think about the preceding analysis procedures (stratifying on the claimed values of the variable we wish to measure, estimating the mean in-sample evaluated value for each stratum, and then taking the sum of these means weighted by the population of sites in each stratum), is that we are fitting an ANOVA-style dummy regression model, where the dependent variable in the regression is evaluated site savings, and the dummy variable regressors represent membership in different strata. That is, a regression model of the form

$$y_k = \sum_{h=1}^{H} \delta_{h,k} \beta_h + \varepsilon_k$$

where $\delta_{h,k}$ is a dummy variable, which is one if the *k*-th site belongs to the *h*-th stratum. The β_h coefficient estimates for these $\delta_{h,k}$ dummy variables are each stratum's sample mean of evaluated savings.

In this regression, claimed savings numbers are used only to determine stratum membership. In effect, evaluated savings for sites not in the sample are estimated simply as the sample mean of evaluated savings in that stratum. This point of view prompts one to wonder if there is not a way to use each site's claimed savings directly to better estimate evaluated savings at non-sampled sites. In fact there is such a method. In its simplest form it is the well-known ratio estimator. That is, we think of the relationship between evaluated savings and claimed savings as being

$$y_k = \beta x_k + \varepsilon_k$$

where y_k is the evaluated savings for site k and x_k is claimed savings for the same site. We can think of the ratio estimator, like the preceding model, as the estimation of a regression relationship; but rather than a dummy variable regression we are now estimating a simple regression without an intercept.

Whereas before our task was to estimate means for each stratum, now our task is to estimate a single ratio coefficient β. Since we are thinking of this as a regression relationship, let us assume that the k-th site error term ε_k has a variance σ_k^2 which is known at least to a factor of proportionality, and which may vary across sites. Assuming that a sample has been drawn, how would one go about estimating the ratio coefficient? One could of course use the ordinary least squares formula, with evaluated savings as the dependent variable and claimed savings as the regressor. But given that we have a sample in which different sites may have had different selection probabilities π_k and given that we are allowing the error term variance σ_k^2 to differ across sites, this is suboptimal. A better approach is to use a formula which includes a GLS weight to deal with differing error variances and a π-weighting to deal with differing selection probabilities:

$$\hat{\beta} = \frac{\sum_s x_k y_k / \pi_k \sigma_k^2}{\sum_s x_k^2 / \pi_k \sigma_k^2}$$

Our estimate for population evaluated savings is then the sum of all claimed savings times this estimated ratio coefficient[7]. A frequently used simplifying assumption is that the error term variances σ_k^2 are proportional to x_k. Substituting this into the above formula leads to the radical simplification:

$$\hat{\beta} = \frac{\sum_s y_k / \pi_k}{\sum_s x_k / \pi_k}$$

This is the classic ratio estimator. We note that it is merely the ratio of the direct, ANOVA-style stratified sample-style population total estimate for evaluated savings to what appears to be that for claimed savings (of course we actually know the values of x_k for the entire population).

[7] Actually, this is not quite accurate. We know from regression theory that the sum of residuals in a fitted GLS regression vanishes if error term variance can be expressed as a linear function of the regressors. This holds true, for example, in a regression with a constant term and standard OLS-assumption homoskedastic residuals; it holds true for the ratio estimator only if the error term variances σ_k^2 are assumed to be proportional to their x_k's. Otherwise, we need to include an estimate of the program population sum of the error term residuals in our estimate of population total savings: $\sum_{k=1}^{n} e_k / \pi_k$, where the e_k's are the regression residuals and the π_k's are the selection probability for each sampled point.

The above ratio analytical approach could be applied to any sample, assuming the sample design (and consequently the π-weights) is known. It could, for example, certainly be applied to a sample drawn using a Neyman allocation with a Dalenius-Hodges stratification; but that particular sample design scheme is no longer optimal. We refer back to the basic principle that we wish to sample with greater relative frequency sites about which our initial "guesses" are less certain.

Since we are now assuming the regression relationship $y_k = \beta x_k + \varepsilon_k$, with the error term variance σ_k^2 known, the appropriate measure of the uncertainty attached to a given site is now the error term variance. It turns out that it is optimal to give each site a selection probability proportional to its σ_k (the standard deviation, not the variance). This is very like the Neyman allocation rule, where we sampled from a stratum proportionally to its intra-sample standard deviation. But unlike the Neyman allocation rule, where all sites in a given stratum in effect shared a common uncertainty measured by the intra-stratum variability of savings, here we allow the error term variance to differ across individual sites. The classic simplest-case ratio estimator, for example, in effect assumes that the variance varies across sites proportionally to x_k.

A selection probability proportional to its error standard deviation must be attached to each site. Since error standard deviations are permitted to vary across sites, it follows that each site in the population may optimally have a different selection probability. One method for drawing a sample which satisfies this requirement is a list-sequential scheme; we go down a list of sites in our population matched with their appropriate selection probabilities π_k and for each site perform some independent random experiment which results in sample inclusion for that site with likelihood π_k. This scheme, though simple to execute as these things go, has the drawback of a random sample size.

Developing a sample design which delivers a fixed sample size and yet permits each site to have a unique selection probability is in fact difficult. In practice, a suggested way to achieve a fixed sample size along with selection probabilities which are near-optimal is stratified sampling, with the population of sites sorted on their error standard deviation values σ_k. The relative frequency of sampling in each stratum is then taken proportional to the *average* σ_k of sites in that stratum.

In the case of the classical ANOVA model, each stratum introduced implies another parameter (a stratum mean) to estimate. Observe that this is not true here, so one can be more liberal in creating strata. A rule for determining optimal stratum boundaries given a fixed number of strata is the equal aggregate-σ rule: the sum of the σ_k's in each stratum should be approximately equal. This is reminiscent of, but different from, the Dalenius-Hodges equal aggregate $\sqrt{f}$ rule.

Similarities, Differences, and Generalizations

For purposes of constructing a sample design, the classical ANOVA model estimates the uncertainty attached to a given site as the standard deviation of savings across sites in its

stratum. This standard deviation can be proxied by the standard deviation of claimed savings in that stratum. By contrast, the ratio estimator requires that (up to a factor of proportionality) the error-term uncertainty associated with each site be specified *a priori*.

A common, convenient, but not necessarily appropriate assumption is that the error term variance is proportional to the size of claimed savings. An offsetting benefit of the ratio model's greater complexity is that it can be said to provide a direct estimate of the "verification ratio", a quantity of great interest in evaluation work. In addition, the ratio model makes better use of known information (claimed savings) in estimating savings for non-sampled sites. But which is a "better" analytical approach is really an empirical question and depends on how well the model assumptions fit the point scatter encountered in the actual finite program population.

Both the stratified-sample ANOVA mean estimation approach and single ratio approach can be seen as members of a larger class of regression estimators. These regression models tie prior information (known for all sites) to evaluated information known only for sampled sites. More complex regression relations could easily be specified. In the case of the ratio estimator one could, for example, posit a separate ratio for different kinds of sites. In the case of the ANOVA model one could add strata (and consequently additional stratum mean estimates) for different equipment types. In practice the relatively small number of sites in a sample place severe constraints on the number of parameters one should introduce. As anyone who has ever estimated regressions knows, trying to estimate too many parameters with a relatively small data set leads to unstable, noisy parameter estimates.

In determining how many sites to sample to achieve a desired precision level, it is important to be conservative, that is, to sample more sites than the formulas suggest are needed. A good reason for this is that the formulas assume we know things we don't in fact know: the Neyman allocation formulas, for example, apply to standard deviations of actual savings, but are used with claimed savings numbers. A prior estimate of needed sample size in the ratio model relies on assumptions regarding the magnitude of error variances.

In general, if the claimed savings numbers turn out to be poor proxies for evaluated savings, the estimates will be noisier than expected. The estimates will still be statistically valid, however. If, in effect, the ratio estimator has a low R-square, or the stratum variances fed into the Neyman Allocation formulas turn out to be way off, the result will be a sample design that, in retrospect, was inaccurate. The statistical calculations performed on this sample will, however, accurately deliver the bad news -- that the variance of our estimates is high.

If we know, or suspect beforehand, that the claimed savings numbers in the program population are very poor proxies for evaluated savings, then rather than simply "riding into the valley of death" it may be appropriate to consider a more complex double-sampling design. The basic idea behind double sampling is to draw an initial large sample for a relatively cursory review - in effect, filling in for the inaccurate program data base; and on the basis of what is learned in this first pass, drawing a subsample from the large initial sample for more detailed evaluation.

A variant of the double-sampling approach recently developed in evaluation work is the so-called "double ratio analysis"[8]. We stress that the burden in additional complexity of double sampling approaches is high, and they should be contemplated only if the program database is thought to be very bad. The issue is not merely extra work, but also precision. Multi-layered designs in effect estimate more parameters. In the case of double ratio analysis, for example, an overall verification ratio is estimated as the product of two separate component verification ratios, each of which is measured with error; where it is not really needed the double ratio approach can thus produce worse estimates than a one-level sample design.

For the same reasons we think that "Russian doll" approaches, in which a small number of sites receive gold-plated analysis such as end-use metering, a larger group of sites get a less expensive form of engineering evaluation, and a yet larger sample get relatively cursory treatment, should be approached with caution. Not only are such sample designs terribly complex, but if the randomness introduced by each level of sampling is correctly taken into account, they are likely to produce noisier estimates than simpler designs.

In our description of sample designs we have implicitly assumed that the sole criterion for choice of sample design is accurate estimation of evaluated savings in the program population, or equivalently, an overall verification ratio. We think this is appropriate. The principal purpose of evaluation is evaluation. The impressive gains in estimation efficiency achievable through a focused sample design---that is, sampling more heavily where we are more uncertain about what we most want to measure---can easily be lost if the sample design is loaded with competing objectives, such as estimating verification ratios for separate equipment types.

That is not to say that we are prohibited from estimating other quantities of interest; in fact we are free to do so. We noted previously that the modeling approach employed and the sample design are in important particulars independent. There can be only one sample design, and one sample, but many different models can be fitted to the same sample. Assuming there is at least one sausage-making machine retrofit site in the sample, one is free to estimate a separate realization rate for sausage machine retrofits. But the sample design should not be modified to improve the accuracy of the estimate of this relatively unimportant realization rate at the expense of the estimate for the overall realization rate.

Suggested References

For exposition of classical stratified sampling and ratio estimators, we suggest Cochran, *Sampling Techniques (*John Wiley, *1977).* A good reference for the later "regression" approach to sampling discussed here is Sarndal, *Model-Assisted Survey Sampling* (Springer-Verlag, 1992).

[8] See "Double Ratio Analysis: Final Report" Report # CIA-93-X01B, September 1993

4 Quality Assurance Guidelines for Estimating Net-To-Gross Ratios Using Participant Self Reports

4.1 Issues Surrounding the Validity and Reliability of Self-Report Techniques

A central intent of utility DSM program evaluations is to identify that portion of the gross load impacts associated with a program-supported measure installation, that would not have been accomplished in the absence of the program. That portion is the net load impacts. In some cases, net load impacts may be estimated directly using regression models. Where it is not possible to use regression models, an alternate approach to estimating the program impact that is due to free ridership and the net-to-gross (NTGR) ratio (defined as one minus the proportion of free ridership) may be required. This approach commonly involves the use of the self-report method, i.e., asking program participants directly whether they would have installed the same thing without the program. This technique must deal with several methodological problems.

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 to do is report on a hypothetical situation. 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 biased answers 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. Bias in responses is commonly thought to stem from two origins. 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 therespondent. Closely related to this is the possibility that the respondents may want to give an answer that they think will be pleasing to the interviewer. The direction of the first bias would be to increase the NTG ratio, and the second would have an unclear effect – up or down, depending on what the respondent thinks the interviewer wants to hear.

The other 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. This type of motivation could result in an artificially low net-to-gross ratio.

Beyond the fact that the situations of interest have occurred in the past and judgments about them involve hypothetical circumstances, they 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 utility influence difficult to pin down. Some efficiency projects may involve reconfiguration of systems rather than simple installations of energy-efficient equipment.

This type of complexity and variation across sites requires thoughtful design of survey instruments. Following is 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.

These should be regarded as recommendations for minimum acceptable standards for the use of self-report methods to estimate net-to-gross ratios. Much of this chapter focuses on self-report methodologies for developing NTGRs for energy efficiency improvements in all sectors regardless of the size of the expected savings and the complexity of the decision making processes. However, in a given year, energy efficiency programs targeted for industrial facilities are likely to achieve a relatively small number of installations with the potential for extremely large energy savings at each site. Residential programs often have a large number of participants in a given year, but the energy savings at each home, and often for the entire residential sector, are small in comparison to savings at non-residential sites. Moreover, large industrial customers have more complex decision making processes than residential customers. As a result, evaluators are significantly less likely to conduct interviews with multiple actors at a single residence or to construct detailed case studies for each customer – methods that are discussed in detail in the following sections. It may not be practical or necessary to employ the more complex techniques (e.g., multiple interviews at the same site, case-specific NTGR development) in all evaluations. Specifically, sections 4.2, 4.5, 4.7, 4.9, and 4.12 are probably more appropriate for customers with large savings and mare complex decision making processes. Of course, evaluators are free to apply the guidelines in these sections even to customers with smaller savings and relatively simple decision making processes.

4.2 Identifying the Correct Respondent

Recruitment procedures for participation in an interview involving self-reported net-to-gross ratios must address the issue of how the correct respondent(s) will be identified. Complexities to be addressed include situations commonly encountered in large commercial and industrial facilities, such as:

1. Different actors have different and complementary pieces of information about the decision to install, e.g., the CEO, CFO, facilities manager, etc.;
2. Decisions are made in locations such as regional or national headquarters that are away from the installation site;
3. Significant capital decision-making power is lodged in commissions, committees, boards, or councils; and

4. There is a need for both a technical decision-maker and a financial decision-maker to be interviewed (and in these cases, how the responses are combined will be important).

An evaluation using self-report methods should employ and document rules and procedures to handle all of these situations in a way that assures that the person(s) with the authority and the knowledge to make the installation decision are interviewed.

4.3 Set-Up Questions

The decisions that the net-to-gross questions are addressing may have occurred as long as two years prior to the interview. Regardless of the magnitude of the savings or the complexity of the decision-making process, questions may be asked about the motivations for making the decisions that were made, as well as the sequence of events surrounding the decision. Sequence and timing are important elements in assessing motivation and program influence on it. Unfortunately, sequence and timing will be difficult for many respondents to recall two years later, which is the standard schedule for first-year load impact evaluation governed by the Protocols. This makes it essential that the interviewer guide the respondent through a process of establishing benchmarks against which to remember the events of interest. Failure to do so could well result in, among other things, the respondent "telescoping" some events of interest to him into the period of interest to the evaluator. Motivations, competing alternatives, and battles lost could recede in memory. Set-up questions that set the mind of the respondent into the train of events that led to the installation, and that establish benchmarks, can minimize these problems. However, one should be careful to avoid wording the set-up questions in such a way so as to bias the response in the desired direction.

Set-up questions should be used at the beginning of the interview, but they can be useful in later stages as well. Respondents to self-report surveys frequently are individuals who participated in program decisions and, therefore, may tend to provide answers ex post that validate their position in those decisions. Such biased responses are more likely to occur when the information sought in questions is abstract, hypothetical, or based on future projections, and are less likely to occur when the information sought is concrete. To the extent that questions prone to bias can incorporate concrete elements, either by set-up questions or by follow-up probes, the results of the interview will be more persuasive.

An evaluation using self-report methods should employ and document a set of questions that adequately establish the set of mind of the respondent to the context and sequence of events that led to decision(s) to adopt a DSM measure or practice, including clearly identified benchmarks in the customer's decision-making process.

4.4 Use of Multiple Measures

Regardless of the magnitude of the savings or the complexity of the decision-making process, one should assume that using multiple questionnaire items (both quantitative and qualitative) to measure one construct is preferable to using only one item, as it is well-documented in the measurement literature that reliability is increased by the use of multiple

items unless some items are uncorrelated with the other items (Blalock, 1970; Crocker & Algina; 1986; Duncan, 1984).

4.5 Use of Multiple Respondents

In situations with relatively large savings and more complex decision-making processes, one should use, to the extent possible, information from more than one person familiar with the decision to install the efficient equipment or adopt energy-conserving practices or procedures (Patten, 1987; Yin, 1994).

4.5.1 Measures of Reliability

The internal consistency of multiple-item scales should not be assumed. Techniques available for testing reliability include: split-half correlations, alternate forms tests, and Cronbach's alpha (Nunnally, 1978; Crocker & Algina, 1986; Cronbach, 1951; DeVellis, 1991).
An evaluation using self-report methods should employ and document some or all of these tests or other suitable tests to evaluate reliability, including a description of why particular tests were used and others were considered inappropriate.

For those sites with relatively large savings and more complex decision-making processes, both quantitative and qualitative data may be collected from a variety of sources (e.g., telephone interviews with the decision maker, telephone interviews with others at the site familiar with the decision to install the efficient equipment, paper and electronic program files, and on-site surveys). These data must eventually be integrated in order to produce a final NTGR.[9] Of course, it is essential that all such sites be evaluated consistently using the same instrument. However, in a situation involving both quantitative and qualitative data, interpretations of the data may vary from one evaluator to another, which means that, in effect, the measurement result may vary. Thus, the central issue here is one of reliability, which can be defined as obtaining consistent results over repeated measurements of the same items.

To guard against such a threat at those sites with relatively large savings and more complex decision-making processes, the data for each site should be evaluated by more than one member of the evaluation team. Next, the resulting NTGRs for the projects should be compared, with the extent of agreement being a preliminary measure of the so-called inter-rater reliability. Any disagreements should be examined and resolved and all procedures for identifying and resolving inconsistencies should be thoroughly described and documented (Sax, 1974; Patton, 1987).

4.5.2 Handling Apparent Inconsistencies

When multiple questionnaire items are used to calculate a free ridership probability there is always the possibility of apparently contradictory answers. Contradictory answers indicate problems of validity and/or reliability (internal consistency). Occasional inconsistencies indicate either that the respondent has misunderstood one or more questions, or is answering according to an unanticipated logic. Apparent inconsistencies should be identified and

[9] For a discussion of the use of qualitative data see Section 4.11.

handled before the interview is over. If the evaluator waits until the interview is over to consider these problems, there may be no chance to correct misunderstandings on the part of the respondent or to detect situations where the evaluator brought incomplete understanding to the crafting of questions. In some cases, the savings at stake may be sufficiently large to warrant a follow-up telephone call to resolve the inconsistency.

However, despite the best efforts of the interviewers, some inconsistencies may remain. When this occurs, evaluator could decide which of the two answers, in their judgement has less error, and discard the other. Or, one could weight the two inconsistent responses in a way that reflects the evaluator's estimate of the error associated with each, i.e., a larger weight could be assigned to the response that, in their judgement, contains less error.

Finally, an evaluation report using self-report methods should describe the approach to identifying and resolving apparent inconsistencies. The report should include : 1) a description of contradictory answers that were identified, 2) whether and how it was determined that the identified inconsistencies were significant enough to indicate problems of validity and/or reliability (internal consistency), and 3) how the indicated problems were mitigated. These rules for resolving inconsistencies should be established, to the extent feasible, before the analysis begins. Details regarding the establishment and use of such rules are provided in Section 4.11.2.

4.5.3 Consistency Checks

One of the potential problems with self-report methods is the possibility of answering the questions in a way that conforms to the perceived wishes of the interviewer, or that shows the respondent in a good light. One of the ways of mitigating these tendencies is to ask one or more questions specifically to check the consistency and plausibility of the answers given to the core questions. Inconsistencies can highlight efforts to "shade" answers in socially desirable directions. While consistency checking won't overcome a deliberate and well-thought-out effort to deceive, it will often help where the process is subtler or where there is just some misunderstanding of a question.

An evaluation using self-report methods should employ a process for setting up checks for inconsistencies when developing the questionnaire items, and describe and document the methods chosen as well as the rationales for using or not using the techniques for mitigating inconsistencies.

4.6 Making the Questions Measure-Specific

It is important for evaluators to tailor the wording of central free ridership questions to the specific technology or measure that is the subject of the question. It is not necessarily essential to incorporate the specific measure into the question, but some distinctions must be made if they would impact the understanding of the question and its potential answers. For instance, when the customer has installed equipment that is efficiency rated so that increments of efficiency are available to the purchaser, asking that respondent to indicate whether he would have installed the same equipment without the program could yield confusing and imprecise answers. The respondent will not necessarily know whether the evaluator means the exact same efficiency, or some other equipment at similar efficiency, or just some other equipment of the same general type. Some other possibilities are:

1. Installations that involve removal more than addition or replacement (e.g., delamping or removal of a second refrigerator or freezer in a residence);
2. Installations that involve increases in productivity rather than direct energy load impacts;
3. Situations where the energy-efficiency aspect of the installation could be confused with a larger installation; and
4. Installation of equipment that will result in energy load impacts, but where the equipment itself is not inherently energy-efficient.

An evaluation using self-report methods should include and document an attempt to identify and mitigate problems associated with survey questions that are not measure-specific, and an explanation of whether and how those distinctions are important to the accuracy of the resulting estimate of freeridership.

4.7 Partial Freeridership

Partial freeridership can occur when, in the absence of the program, the participant would have installed something more efficient than the program-assumed baseline efficiency but not as efficient as the item actually installed as a result of the program. When there is a likelihood that this is occurring, an evaluation using self report methods should include and document attempts to identify and quantify the effects of such situations on net savings. Partial free-ridership should be explored for those customers with large savings and complex decision making processes.
In such a situation, it is essential to develop appropriate and credible information to establish precisely the participant's alternative choice. The likelihood that the participant would really have chosen a higher efficiency option is directly related to their ability to clearly describe that option.

An evaluation using self-report methods should include and document attempts to identify and mitigate problems associated with partial freeridership, when applicable.

4.8 Deferred Freeridership

Deferred free riders are those customers who would, in the absence of the program, have installed exactly the same equipment that they installed through the utility DSM program, but the utility induced them to install the equipment earlier than they would have otherwise. That is, the utility *accelerated* the timing installation of the equipment. Because determining the extent of utility influence on the timing of the installation is a complex process, an evaluator should avoid relying on a single question asked of the key decision-maker. Rather, an evaluator should examine all available data and determine whether the preponderance of evidence supports the conclusion of deferred free ridership. Data from such sources as additional closed- and open-ended questions asked of the key decision-maker, information obtained from other people at the site familiar with the decision to install the efficient equipment, and information gathered from the program paper files should also be collected and analyzed. Rules for integrating the responses to closed- and open-ended questions should

be established, to the extent feasible, before the analysis begins. Details regarding the establishment and use of such rules are provided in Section 4.11.2.

Unfortunately, evaluation budgets may only permit such data to be collected and analyzed for those customers with larger savings. For those customers with the smaller savings, the NTGR may be based only on the responses from close-ended questions obtained from the key decision-maker. In such cases, closed-ended questions regarding utility influence on both *what* was installed and *when* it was installed could be asked. These answers could be analyzed mechanically using an algorithm. However, to the extent that closed-ended questions are unable to capture fully the complexity of the decision-making process, any resulting conclusions regarding deferred free ridership may be biased, with the direction of the bias unknown.

Whenever deferred free ridership is identified by a utility, the onus is on the utility to account for such free ridership in the stream of future utility savings. This could be done by calculating a *lifecycle* NTGR and applying it throughout the effective useful life of the equipment. Or, a utility could calculate a *first-year* NTGR and adjust the stream of savings to account for the fact that the savings associated with deferred free riders will be reduced to zero in the year in which they said they would have installed the same equipment in the absence of the program.

4.9 Third-Party Influence

Currently, there is no standard method for capturing the influence of third parties on customer's decision to purchase energy efficient equipment. Third parties who may have influence in this context include market actors such as store clerks, manufacturers (through promotional literature, demonstrations, and in-person marketing by sales staff), equipment distributors, installers, developers, engineers, energy consultants, and architects. When one chooses to measure the effect of third parties, one should keep the following principles in mind:

1. The method chosen should be balanced. That is, the method should allow for the possibility that the third-party influence can increase or decrease the NTGR that is based on the customer's self report.

2. The rules for deciding which customers will be examined for potential third party influence should balanced. That is, the pool of customers selected for such examination should not be biased towards ones for whom the evaluator believes the third-party influence will have the effect of influencing the NTGR in only one direction.

3. The plan for capturing third-party influence should be based on a well-conceived causal framework.

The onus is on the evaluator to build a compelling case using a variety of quantitative and/or qualitative data for changing the customer's NTGR.

4.10 Scoring Algorithms

A consequence of using multiple questionnaire items to assess the probability of freeridership (or its complement) is that decisions must be made about how to combine them. Should two items be averaged or should one supersede the other? Do all items have equal weight or are some more important indicators than others? Answers to these questions can have a profound effect on the final NTGR estimate. These decisions are incorporated into the algorithm used to combine all pieces of information to form a final result. All such decisions must be described and justified by evaluators.

4.11 Handling Non-Responses and "Don't Knows"

In this section, we address the situation where customers selected for the evaluation sample refuse to be interviewed or do not complete an attempted interview or questionnaire. When this happens, a decision must be made about how to treat that case in the process of aggregating participant-level results to program-level results. For example, making no decision assumes that the non-respondents would have answered the questions at the mean. Thus, their net-to-gross ratios would assume the mean value. This may or may not be a reasonable assumption, but it should not go unexplained. It is essential to do an analysis to determine the characteristics of the non-respondents in order to decide what assumptions should be made about their unanswered questions. Evaluators should do such an analysis and make judgments on what customer characteristics are likely to be relevant to answers to net-to-gross questions. These judgments and the decisions that flow from them must be described and rationales provided for them.

Respondents who do answer interview questions may nevertheless answer some questions with a "don't know" response. When this answer is received for a question included in the net-to-gross algorithm, decisions must be made about how to handle such a response. It is clear that some questions are more central than others, implying different assumptions and decisions. A decision about a "don't know" answer to core questions concerning what the respondent would have done absent the program may well be different than the decision about handling the "don't know" answer to a question about the timing of learning about the program or the timing of measure installation without the program. Evaluators should decide, in advance, how to handle "don't know" answers and justify those decisions.

4.12 The Use of Qualitative Data and Reporting Requirements

The M&E Protocols focus entirely on quantitative methods that stress such elements as quasi-experiments, paper and pencil "objective" instruments containing closed-ended questions, and multivariate statistical analyses. However, many DSM 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 a *quantitatively*-based results (Britan, 1978; Weiss and Rein, 1972; Patton, 1987). *Qualitative* methods stress in-depth, open-ended interviews, direct observation, and written documents, including such sources as open-ended questions and program records.

There is wide agreement on the value of *both* qualitative and quantitative data in the evaluation of many kinds of programs. Moreover, it is inappropriate to cast either approach

in an inferior position. The complexity of organizational decisions regarding the purchase of efficient equipment can be daunting, especially in large organizations for which the savings are often among the largest. In such situations, the reliance on only quantitative data can miss some important elements of the decision. The collection and interpretation of qualitative data can be especially useful in broadening our understanding of a utility's role in this decision.

When one chooses to complement a quantitative analysis of the NTGR with a qualitative analysis, there are a few very basic concerns that one must keep in mind.

4.12.1 Data Collection

Information relevant to the purchase and installation decision can include:

1. Program paper files (correspondence between DSM program staff and the customer, evidence of economic feasibility studies conducted by the utility or the customer, correspondence among the customer staff, other competing capital investments planned by the customer)
2. Program electronic files (e.g., program tracking system data, past program participation)
3. Interviews with other people at the site who are familiar with the program and the choice (e.g., operations staff)
4. Open-ended questions on structured interviews with the key decision maker and other staff who may have been involved with the decision.

Where appropriate, for example, in the case of large-scale commercial and industrial sites, these data should be organized and analyzed in the form of a case study.

4.12.2 Establishing Rules for Data Integration

Before the analysis begins, one should establish, to the extent feasible, rules for the integration of the quantitative and qualitative data. 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 freeridership.

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 one's judgement in deciding how much weight to given to the quantitative and qualitative data and how to integrate the two.

4.12.3 Analysis

A case study is an organized presentation of all the information available about a particular customer site with respect to all relevant aspects of the decision to install the efficient equipment. When a case study approach is used, the first step is to pull together the data relevant to each case and write a discrete, holistic report on it (the case study). In preparing the case study, redundancies are sorted out, and information is organized topically. *This information should be contained in the final report.*

The next step is to conduct a content analysis of these data. This involves identifying coherent and important examples, themes, and patterns in the data. The analyst looks for quotations or observations that go together and that are relevant to the *customer's decision to install the efficient equipment*. Guba (1978) calls this process of figuring out what goes together "convergence," i.e., the extent to which the data hold together or dovetail in a meaningful way. Of course, the focus here is on evidence related to the degree of utility influence in installing the efficient equipment.

Sometimes, *all* the 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.

Finally, it must be recognized that there is no single right way to conduct qualitative data analysis:

> The analysis of qualitative data is a creative process. There are no formulas, as in statistics. It is a process demanding intellectual rigor and a great deal of hard, thoughtful work. Because different people manage their creativity, intellectual endeavors, and hard work in different ways, there is no one right way to go about organizing, analyzing, and interpreting qualitative data. (p. 146)

Ultimately, if the data are systematically collected and presented in a well-organized manner, and if the arguments are clearly presented, any independent reviewer can understand and judge the data and the logic underlying any NTGR. Equally important, the independent reviewers will have all the essential data to enable them to replicate the results, and if necessary, to derive their own estimates.

4.13 Weighting

The Protocols require estimates of the NTGR at the end use and program levels. Of course, such an NTGR must take into account the size of the impacts at the customer or project level. Consider two large industrial sites with the following characteristics. The first involves a customer whose self-reported NTGR is .9 and whose estimated annual savings are 200,000

kWh. The second involves a customer whose self-reported NTGR is .15 and whose estimated savings are 1,000,000 kWh. One could calculate an unweighted NTGR across both customers of .53. Or, one could calculate a weighted NTGR of .28. Clearly, the latter calculation is required.

It is critical to recognize that how these NTGRs are applied by utilities in order to estimate the stream of benefits and earnings can produce very different results. First note that in order to produce a single end-use NTGR for all fuel impacts, one must first convert both gross and net kWh, kW, and therm impacts into a common metric, dollars.[10] Once converted, the total end-use net impacts can be divided by the total end-use gross impacts to produce an end-use NTGR. Now, suppose a utility calculates an end-use NTGR and applies it to the stream of kWh, kW, and therm impact estimates. When this is done, certain distortions can occur. For example, if a customer has relatively small kWh and kW impacts but enormous therm impacts, the therm impacts will dominate the end-use NTGR. If a utility, in calculating its earnings claim, applies this end-use NTGR to the separate benefits streams for kWh, kW, and therms contained in its "E" tables, the net *kWh and kW* impacts will be inflated. The appropriate approach is to calculate three separate NTGRs for kWh, kW, and therms within each end use. A utility could then apply these NTGRs to the separate benefits streams for kWh, kW, and therms in their "E" tables.

4.14 Assessing Spillover

Spillover is defined as :

> Reductions in energy consumption and/or demand in a utility's service area caused by the presence of the DSM program, beyond program-related gross savings of participants. These effects could result from: (a) additional energy efficiency actions that program participants take outside the program as a result of having participated; (b) changes in the array of energy-using equipment that manufacturers, dealers, and contractors offer all customers as a result of program availability; and (c) changes in the energy use of non-participants as a result of utility programs, whether direct (e.g., utility program advertising) or indirect (e.g., stocking practices such as (b) above, or changes in consumer buying habits).[11]

Part "a" of above definition is referred to as *participant spillover*. *The following recommendations apply only to estimating participant spillover.*

All of the measurement issues that have been identified in these guidelines for attributing installations of energy-efficient equipment to program influence apply to spillover installations as well. It is important to remember that evaluations that include savings from spillover measures must estimate the gross savings using the same level of methodological rigor that was used for program-induced measures and practices. In addition, there are extra

[10] This can be done using the marginal costs associated with various costing periods.

[11] *Protocols and Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management Programs*, page A-9.

hurdles that evaluators must address if a persuasive case is to be made for program influence on these installations. These hurdles stem from the fact that the identification of appropriate installations and their connection to a utility program is necessarily more vague (less concrete) than was the case for equipment specifically recommended or rebated by utilities. The reason is obvious. In traditional program evaluations, specific equipment is specified in program records, serving both the identification and the program connection functions. The issue is only in assessing the *level* of program impact on the installation decision at some point between 0 and 100 percent. For spillover measures, simply identifying the equipment and/or practices is at issue, as well as making any connection at all with a utility program. Evaluations that include spillover measures in net program impact should specify how each of the issues identified in this and previous sections have been addressed. Some acknowledgment of the "softness" of simple statements of utility influence, coupled with specific efforts to strengthen confidence in the statements, should be included in the evaluation report.

There are many issues surrounding the matter of defining what equipment and/or practices are appropriate for consideration in spillover analyses. These are beyond the scope of these guidelines. Criteria for what is appropriate will vary by the utility, and are equally at issue for self-report as for other methods of assessing impact. The process of eliciting, from the respondent, what installations, modifications, and reconfigurations have been completed within a specified time period is important, but not subject to these guidelines that are oriented only to self-reported program influence. It may be important, however, to state that identified spillover measures and/or practices must be separated from measures and/or practices that have been claimed for direct program influence in other evaluations. To avoid double counting, this will probably require that the respondent be informed of the installations listed in program tracking systems that have been claimed in other evaluations for direct program influence to avoid double counting.

When installations have been identified as potential spillover cases, the respondent must be asked about the level of utility influence on the selection of the energy-efficient version of the equipment installed, modified, or reconfigured. Because the evaluator has eliminated from consideration all equipment directly associated with the utility's programs, any influence identified by the respondent will usually be indirect, i.e., less than concrete. It therefore becomes important to assess the credibility of any claims the respondent makes for utility influence on the decision. Usually this will mean asking questions that attempt to tie these "soft" claims to something more concrete. This might mean establishing the means by which the influence occurred, identifying third parties involved in the communication of information or in the influence of attitudes, and indicating a time period and context in which the influence took place. The more concrete the ties, the more persuasive the case for claimed influence will be.

Appendix A: References

ASHRAE, "Energy Estimating Methods." Chapter 28 in *The 1993 ASHRAE Handbook: Fundamentals*. Published by the American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Inc., 1993.

Berk, R. A. "A Primer on Robust Estimation." In *Modern Methods of Data Analysis*, edited by Fox, J. and Long, J. S., Newbury Park, CA: Sage Publications, 1990.

Belsey, D. A., E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley. 1980.

Blalock, H. (1970), "Estimating Measurement Error Using Multiple Indicators and Several Points in Time," *American Sociological Review,* 35, pp. 101-111.

Bogdan, Robert and Steven J. Taylor. Introduction to Qualitative Research Methods. New York: John Wiley & Sons, 1975.

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

Campbell, Donald T. and Julian C. Stanley. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally College Publishing, 1963.

Cochran, William G. *Sampling Techniques*. New York: John Wiley & Sons, 1977.

Cohen, Jacob and Patricia Cohen. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. New York: John Wiley & Sons, 1975.

Cook, Thomas D. and Donald T Campbell. *Quasi-Experimentation: Design & Analysis Issues for Field Settings.* Boston, MA.: Houghton Mifflin Company, 1979.

Crocker, L. & Algina, J. (1986) *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart & Winston.

Cronbach L.J. (1951). "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika*, 16, 297-334.

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

Dubin, J., and D. Rivers. "Selection Bias in Linear Regression, Logit and Probit Models." Chapter 10, *Modern Methods of Data Analysis*. California: Sage Publications, 1990.

Duncan, O.D. (1984). Notes on Social Measurement: Historical and Critical. New York: Russell Sage.

Electric Power Research Institute (a). *Impact Evaluation of Demand-Side Management Programs: A Guide to Current Practice*. EPRI CU-7179, Volume 1. 1991.

Electric Power Research Institute (b). *Impact Evaluation of Demand-Side Management Programs: Case Studies and Applications*. EPRI CU-7179, Volume 2. 1991.

Electric Power Research Institute (c). *Impact Evaluation of Demand-Side Management Programs*, (Volumes 1 & 2), EPRI CU-7179s, 1991.

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, 1978.

Johnston, J. *Econometric Methods*. New York: McGraw-Hill Book Company, 1984

Kerlinger, F. N., and E. J. Pedhazur. *Multiple Regression in Behavioral Research*. New York: Holt, Rinehart, and Winston. 1973.

Kennedy, Peter. *A Guide to Econometrics*. Cambridge, MA: The MIT Press, 1992

Kish, Leslie. *Survey Sampling*. New York: John Wiley & Sons, 1965.

Kraemer, Helena Chmura and Sue Thiemann. *How Many Subjects*: *Statistical Power Analysis in Research*. Newbury Park, CA: Sage Publications, 1987.

Lipsey, Mark W. *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, CA: Sage Publications, 1990.

Madow, William G., Harold Nisselson, Ingram Olkin. *Incomplete Data in Sample Surveys*.. New York: Academic Press, 1983

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

Parti, C. and M. Parti. "The Total and Appliance-Specific Conditional Demand Analysis for Electricity in the Household Sector." Bell Journal of Economics, Spring 1980.

Pollard, W.E. *Bayesian Statistics for Evaluation Research*. California: Sage Publications. 1986.

Ridge, Richard, Kirtida Parikh, Dan Violette, Don Dohrman, and Katherine Randazzo.

"An evaluation of Statistical and Engineering Models for Estimating Gross Energy Impacts." A report sponsored by the CADMAC Subcommittee on Modeling Standards for End-Use Consumption and Load Impact Models, 1994.

Rossi, Peter and Howard E. Freeman. *Evaluation: A Systematic Approach.* Newbury Park, California: SAGE Publications, 1989.

Sayrs, Lois. *Pooled Time Series Analysis*. Newbury Park, CA: SAGE Publications, 1989.

Sax, Gilbert. Principles of Educational Measurement and Evaluation. Belomont, CA: Wadsworth Publishing Company, Inc., 1974.

SCE(c), *A Review and Critique of Statistical Techniques for Estimating Net Impacts, Volume 8*, 1993

Violette, D., and M. T. Ozog. "Correction for Self-Selection Bias in the Estimation of Audit Program Impacts." In *Proceedings of the ACEEE 1990 Summer Study in Energy Efficiency in Buildings*, American Council for and Energy Efficient Economy, Vol. 6, 1990.

Weiss, R. S. and M.Rein. 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.

Yin, Robert K. Case Study Research: Design and Methods. Newbury Park, California: SAGE Publications, 1994.