Appendix A: Evaluation Designs and Sampling Requirements.

MDNR proposes that utilities be allowed to select evaluators of their choosing, while the Commission should create a set of standards governing evaluation designs and reports and monitor compliance with the standards. MDNR also maintains that the MEEIA rules regarding evaluation (see Staff 2010, Section 8) should be sufficiently flexible to accommodate a range of program designs including some of the examples discussed below. This appendix describes five generic program designs and an appropriate evaluation approach to each. The appendix concludes by describing MDNR's position on the role of probability samples in energy evaluation projects.

A Typology of Evaluations

Five broad types of evaluation designs are summarized in Table A-1. This summary is not intended to be exhaustive. The summary highlights some key points about the scope, design, sampling and appropriate outcomes for each type of design. Individual DSM programs have specific requirements and any design should be constructed to meet a program's unique needs.

Table A-1: Types of Evaluation Designs

Aggregate-level Studies	
Potential Studies	
Scope of Study	General assessment of measure saturation or customer opinions.
Primary Design	Single group, single observation design.
Sampling and Assignment	Should be conducted using a probability sample of customer groups or catchment area. Characteristics of the sample must reflect key characteristics of the catchment area. Differences between sample strata and study groups must be accounted for statistically.
Evaluation Goals	Basic data collection and data reconnaissance
Promotional/Education program studies	
Scope of Study	General assessment of consumer awareness of utility DSM activities.
Primary Design	Single group, single observation design.
Sampling and Assignment	Should be conducted using a probability sample of customer groups or catchment area. Characteristics of the sample must reflect key characteristics of the catchment area. Differences between sample strata and study groups must be accounted for statistically.
Evaluation Goals	Basic data collection and data reconnaissance
Market transformation program studies	
Scope of Study	Market-wide assessment of changes in the saturation of energy efficiency measures and the adoption of energy efficiency practices
Primary Design	Baseline comparison study, comparison of pre-study and post-study assessments of market conditions or market actors
Sampling and Assignment	Aggregate assessment of measure penetration at two points in time. Background characteristics of the study groups need to be equivalent in order to claim that saturation differences are due to changes in the market place.
Evaluation Goals	Document changes in market, saturation of measures
Individual-level Studies	
Pilot program studies	
Scope of Study	Small-scale assessment of individual measures, test of innovative procedures.
Primary Design	Random Controlled Trial (RCT), a pre-selected group of participants are randomly assigned to study or control group(s) to facilitate the assessment of a program's impact.
Sampling and Assignment	Random assignment to study group. Insures that groups are equivalent at baseline and that any energy usage changes are attributable to measure rather than other factors.
Evaluation Goals	Demonstrate the effectiveness of particular measures and programs
Ongoing program studies	
Scope of Study	Large to mid-size study of program participation and energy usage patterns within a geographic area.
Primary Design	Quasi experimental design. A general survey of an area, end-user sector or customer class. The program participation variable is allowed to vary naturally.
Sampling and Assignment	A random sample of end users, identification of program participants through a survey or though use of utility program data (e.g., through registration or rebate forms). Relies on naturally occurring groups of users.
Evaluation Goals	Determining the appropriate attribution of savings, identifying evidence-based measures of free-ridership and spillover. Validate and update measures and measure savings

MDNR's approach to types of evaluations considers the type of empirical comparison at the root of an evaluation study. This divides the five pure evaluation types listed in Table A-1 into two broad categories, those that collect and describe information about aggregate groups of customers, etc., and those that collect and describe information about individuals.

Aggregate-level Studies

Aggregate-level studies consider a utility's end users or market space as a whole. Subgroups may exist within the aggregate population, and a well designed study will adjust its sampling methodology to account for the groups, but the focus of the design and study is to summarize and describe the characteristics of the aggregate, rather than describing the actions of individuals.

Such studies measure characteristics such as aggregate energy savings potential, measure saturation, the reach of educational and promotional materials, and the overall impact of utility DSM programs on the marketplace. This last type of study, the Market Transformation study, is unique because it requires two assessments of a market area to assess differences in saturation occurring during the term of a utility program (see Neji, 2001 for a description of Market Transformation studies).

Because these studies are designed to describe aggregates, e.g., entire user groups, market areas, etc., many questions about the proper attribution of DSM program participation (i.e., identification of free-riders, spillover, and the estimation of net-to-gross ratios) are not appropriate. Identifying and classifying program participants require individual-level measurement and analysis, techniques that are not practical when assessing the state of an entire marketplace or catchment area.

In aggregate-level studies, the characteristics of the sample group must reflect the characteristics of the underlying population. Demonstrating that a sample represents the population supports claims of measure accuracy. Additionally aggregate samples need to be investigated to identify between-group differences and, where necessary, these group differences need to be reflected in any final estimate. Testing for group differences requires a probability sample. For example, consider a sample where awareness of a promotional campaign varies according to the age category of respondents. With a probability sample it is possible to test whether these group differences are statistically significant (e.g., using an analysis of variance or a regression model) and, if they are statistically significant, account for these differences in the description of the effectiveness of the campaign. On the other hand, with a non-probability sample, one cannot determine whether differences are sufficiently large to

require adjustment of an estimate because there is no way to estimate an unbiased variance. This issue is discussed in the section on sampling.

Individual-level Studies

The other major category of evaluation studies focuses on the energy savings and behavior of individual users. Individual-level studies differ from aggregate-level studies primarily in their focus. Instead of making large holistic descriptions of an entire service area, individual-level studies focus on end-user decision making. Assessing the prevalence of free-riders among the participants in a program is possible, as is the relatively accurate estimation of net savings.

Two general types of studies are included in this category: pilot studies of measures and ongoing program studies. Pilot studies are typically small-scale tests of the potential savings possible from new and untested measures. The relatively small size of the study groups allows the investigator to assign participants to groups in a manner that produces equivalent groups at the beginning of the study. This allows the investigator to attribute any differences in energy use observed at the end of the study to the measures assessed.

By providing the evaluator the ability to control assignment and other conditions, evaluations of pilot studies can resemble randomly controlled trials (RCT). Randomly assigning participants to conditions allows the evaluator to attribute any observed differences between groups at the end of the study to the measures being tested.

With “ongoing program studies” the evaluator does not have the same level of control over the assignment of participants to conditions as found in pilot studies. Rather, participants are identified by some methodology, and then further classified into free-riders, etc., using survey questions. These types of studies are called “quasi-experimental designs” because the basis for comparison is based on observed differences between participants, rather than differences an evaluator can control. Analysis of quasi-experiments requires an extensive use of inferential statistics for partitioning savings amounts among different types of users (see Kandel’s description of two-stage linear modules in the determination of net-to-gross measures, 2002) and for controlling for initial differences using analysis of covariance.

This very brief description of different types of evaluation designs has focused on unique types of studies. In practice, types of studies are used in combination to meet the requirements of individual programs.

Estimating energy savings: Dealing with samples

The designs above rely on construction of a representative sample of customers in order to make appropriate inferences about the population of customers participating in a program. The discussion of evaluation samples begins with the simple observation that all measures of energy savings are estimates.

For example, while different pieces of equipment may consume electricity at different rates, the ways a piece of equipment is actually used influences a measure's total energy consumption and, where appropriate, its energy savings. In many cases it is difficult to accurately predict patterns of use, so it is difficult to accurately assess the actual savings associated with a piece of equipment or a practice. In practical terms, an estimate of savings is all that is possible.

In creating these estimates of savings, one uses the available information to describe the circumstances of the, typically, unknown whole. That is, one is using the information collected from a group of observed elements to infer something about a larger group of unobserved elements. The activity of estimating energy savings among a population of un-measured elements, whether they are customers, buildings, offices, etc., from a sample of measured elements is a central concern of inferential statistics. The summary of evaluation methods presented here is informed by this concern.

The task of estimating the characteristics of a population from a sample requires documentation of the probabilities of selection from the population and measuring the variance associated with a measure. These probabilities account for the uncertainty in an estimate introduced by not collecting the measurement from every element in a population. The variance statistic measures the average differences in a set of observations, relative to measures of central tendency (e.g., means and percentages).

The majority of values presented in an evaluation report are measures of central tendency. These values will differ among and between groups for one of two reasons: either because of differences between the observed elements or because of the uncertainty introduced by the sampling procedure. An analyst tries to control the uncertainty of an estimate by controlling the sampling procedure.

Measures of statistical variance, standard deviations and standard errors, are central to statistical inference. These quantities are in the denominator of virtually every test statistic used to assess differences between groups. Having an unbiased estimator of sample variance, i.e., unbiased by the uncertainty introduced by sampling, is essential to making

the correct decision about whether observed differences in a measure of central tendency are due to actual differences in the population or are due to the uncertainty introduced by sampling.

Controlling for the uncertainty introduced by a sampling technique has led to the development of statistical procedures that produce unbiased estimates of sample variance (Lohr, 1999: 2-8). Probability sampling is the only method proven to provide unbiased estimates. Other types of methods, i.e., non-probability samples (such as quota samples) will introduce biases into the data. Using a probability sample allows an analyst to partition the variance around a measure of central tendency into components that reflect the observed differences between elements and components due to sample uncertainty. It is not possible to do this with a non-probability sample because one cannot specify a sampled element's probability of being selected from the population. Results from a non-probability sample cannot be tested using methods of statistical inference because it is not possible to measure and control for sample uncertainty. For example, analyses based on quota samples typically do not consider differences between groups because it is not possible to determine how much of an observed difference is due to the uncertainty introduced by the sampling procedures.

The description of the evaluation types below presumes that all samples are drawn using a probability sample. Relying on non-probability samples produces inconsequential estimates, estimates that cannot be verified and must be regarded as biased.

The final point to be made is about identifying sample bias. Simply defined, sample bias is the difference between the value of a measure collected from a sample and the value of that measure in the population (see Kish, 1965: 11-13). Because the value of the measure in the population is unknown, one has to assess the existence of sample bias through a series of statistical tests (i.e., one cannot determine bias by calculating the difference between an observed quantity from a sample and the corresponding quantity from the population). These tests are necessarily negative. There is no statistical test that will determine that a bias exists. Rather an analyst has to identify the possible sources of sample bias (i.e., due to non-representation of the sample, due to sample non-response, etc.) and eliminate them before determining that the measure taken from a sample is representative of the population. If an analyst cannot conduct these tests, e.g., because of the use of a non-probability sample, one cannot conclude that a sample is representative of the population. The most one can conclude is that the results of study are undetermined, that it is not possible to assert that a result is unbiased. Because of the undetermined nature of results drawn from non-probability samples, the use of probability samples is preferred.

References

Kandel, A. V. (2002) *Theory-Based Estimation of Energy Savings from DSM, Spillover, and Market Transformation Programs Using Survey and Billing Data.* Retrieved October 14, 2009 from http://www.energy.ca.gov/papers/2002-08-18_aceee_presentations/PANEL-10_KANDEL.PDF

Kish, L. (1965) *Survey Sampling*. New York: John Wiley and Sons.

Lohr, S. L. (1999) *Sampling*: Design and Analysis. Pacific Grove, California: Duxbury.

Missouri Public Services Commission (2010) *Demand Side Program Investment Rule*. Docket EW-2010-0265. April 8, 2010.

Neji, L. (2001) "Methods for evaluating market transformation programmes: experience in Sweden." Energy Policy 29 (2001): 67-79.