Reasonable Doubts: Monitoring and Verification for Performance Contracting

Miriam L. Goldberg, XENERGY, Inc.

Performance contracting is expanding as a means of achieving energy efficiency in the private and public sectors. A central issue in negotiating and implementing the performance contract is monitoring and verification requirements. These requirements must balance the need to assure that the contracted savings have been achieved against the costs of high levels of assurance.

Recent developments in Monitoring and Verification (M&V) protocols include the 1993 NAESCO Standards, the Department of Energy's ongoing National Energy Measurement and Verification Protocol (NEMVP), and M&V manuals for individual utilities. These protocols specify the points to be monitored and engineering calculations to be performed for different technologies, but devote limited attention to sampling. In practice, however, sampling requirements often become a major contractual issue, and represent the key to both the accuracy and cost of verification.

In this paper, we discuss approaches to developing cost-effective monitoring and verification protocols, with an emphasis on sampling plans. We review general issues and objectives of monitoring and sampling. We then review the sampling rules for some existing M&V protocols, in relation to those issues. Finally, we present alternate approaches to defining and meeting objectives, with specific examples from current programs.

The discussion is not limited to statistical precision based on ideal implementation of an optimal sample design. Rather, the goal is to distinguish both the statistical and nonstatistical concerns that motivate certain approaches, so that procedures can be developed that provide satisfactory assurances at a minimal total cost.

INTRODUCTION

Performance contracting is expanding as a means of achieving energy efficiency in the private and public sectors. A central issue in negotiating and implementing the performance contract is monitoring and verification requirements. These requirements must balance the need to assure that the contracted savings have been achieved against the costs of high levels of assurance.

Recent developments in Monitoring and Verification (M&V) protocols include the National Association of Energy Service Company's Standards (NAESCO 1993), the Department of Energy's ongoing National Energy Measurement and Verification Protocol (NEMVP 1995), and M&V manuals for individual utilities. These protocols specify the points to be monitored and engineering calculations to be performed for different technologies, but devote limited attention to sampling. In practice, however, sampling requirements often become a major contractual issue, and represent the key to both the accuracy and cost of verification.

In this paper, we discuss approaches to developing cost-effective monitoring and verification protocols, with an emphasis on sampling plans. We review general issues and objectives of monitoring and sampling. We then review the sampling rules for some existing M&V protocols, in relation to those issues. Finally, we present alternate approaches to defining and meeting objectives, with specific examples from current programs.

The discussion is not limited to statistical precision based on ideal implementation of an optimal sample design. Rather, the goal is to distinguish both the statistical and nonstatistical concerns that motivate certain approaches, so that procedures can be developed that provide satisfactory assurances at a minimal total cost.

ISSUES IN SAMPLING FOR M&V

Why Monitor?

Monitoring for performance contracts can serve several purposes. The importance of monitoring and the associated accuracy requirements depend on the importance of these different objectives.

(1) The most basic reason to monitor in this context is to determine the total value of savings from the installed efficiency measure, as a basis for determining payments to be made.

(2) Monitoring may also be seen not simply as a way of confirming what has happened, but as a quality assurance tool to induce a higher level of performance. That is, it may be felt that installations will be more consistent and of higher quality if a higher level of monitoring is imposed.

(3) Finally, monitoring can provide information for future decisions. Beyond the information required to determine payments, monitoring provides information on the value of investments, which can be useful to both purchasers and providers of energy services.

If monitoring is seen primarily as an inducement to higher quality (reason 2), statistical rigor may be less important than the perceived effect of the monitoring on the quality of installations. In principle, a performance-based payment with achieved savings measured with acceptable accuracy (reason 1) should be sufficient inducement to high performance. However, a purchaser of energy services may feel uncomfortable with a sampling scheme that provides sufficient accuracy for accounting purposes but leaves many units unmonitored.

There may be several reasons for this discomfort. One may be a lack of intuitive acceptance of a statistical estimate based on random selection. Regardless of the accuracy bounds reported and the sampling scheme used, the purchaser retains the concern that only a small fraction of units were actually monitored, and these may not be representative of the whole. Another reason for discomfort may be legitimate concerns about potential biases in the sample implementation. In practice, monitoring is constrained by site logistics such as wiring and equipment configurations, and purely random samples can never be achieved in the field. Moreover, if the units to be monitored are identified at the time of installation, or are known for an installation requiring ongoing maintenance by the performance contractor, the monitored units may receive better attention than the nonmonitored ones. This preferential treatment can happen without any conscious desire to thwart the system.

For example, suppose a lighting retrofit is implemented at a large facilities with many buildings of different sizes. To determine the value of savings for the facility as a whole within acceptable accuracy bounds, monitoring the smallest buildings might be unnecessary. However, the purchaser may want some monitoring done in each building as a way of ensuring that all areas receive equal quality attention.

If monitoring is to provide information for future decisions (reason 3) a higher level of monitoring may be appropriate than would be necessary for purposes of establishing a basis for payment. Better accuracy may be needed for the project as a whole, or for particular components.

In most applications, all of these reasons for monitoring will operate to some extent. This paper focuses on monitoring for determination of savings for payment purposes. However, in establishing monitoring requirements for a project, the importance of the other objectives must not be overlooked. Indeed, distinguishing among these objectives can be important in resolving conflicts over the appropriate level of monitoring.

Why Sample?

In many M&V situations, no sampling is required. Some contracts do not require monitoring at all, only verification that implementation has occurred. Verification typically involves observation and counts of installed equipment, with no sampling. In other cases, monitoring is required, but only one unit or a small number of units are involved, and all are to be monitored. An example would be a chiller replacement project. Finally, there are situations where the cost of obtaining data for all units is low enough that there is little value to sampling. For example, in a project to treat a large number of residential units, PRISM (Fels et al. 1995) is the method recommended by the NEMVP for measuring savings. Since the only data this tool requires are existing monthly billing records, there would be almost no cost savings associated with examining only a sample of units rather than all of them.

Sampling is typically used when monitoring is required for a large number of units, and the cost of monitoring all the units is too high to be justified. Sampling may also be used as part of the verification process, for example, if all units are counted but a more detailed inspection is made of every tenth unit. This paper focuses on cases where sampling is required.

How Much to Invest in M&V

Monitoring and verification cost money. The purchaser of energy services wants to be assured that the savings paid for actually occur. However, that assurance is paid for out of those savings. Regardless of how the M&V costs are covered according to the contract structure, money spent on monitoring and verification is money that could otherwise be spent on more or better energy efficiency measures. This consideration applies as much to determining the method and duration of monitoring for a single large measure as it does to determining the number of observations to use in a sampling situation. What distinguishes the sampling case is the ability to calculate accuracy (or the reduction in uncertainty) as a continuous function of the sample size, which is roughly proportional to the cost of monitoring.

Guidelines. There are no hard rules about how to specify the accuracy level required for performance contracting monitoring. The NEMVP (p. 15) indicates that the cost of measurement should not exceed 20 percent of the anticipated net benefit. Similar rules of thumb have been used in the context of DSM program evaluation. Just as in the evaluation case, more data collection may be warranted if the information has value beyond the immediate needs of the contract; conversely, less data collection may suffice to give high confidence in a well established technique with a trusted implementer, or where unit costs for data collection and analysis are low.

The NEMVP offers as an example (p. 22) a project with an anticipated net benefit of $100,000/year, with initial bounds of ± $20,000/year. The protocols suggest that it may be reasonable to spend $10,000/year to reduce the bounds to ± $10,000/year, but would not be worth $30,000. Certainly it is unreasonable to spend $30,000 to eliminate $10,000 worth of uncertainty, unless there is some other benefit of the information that justifies the investment. In most cases, however, it would not even be worth $10,000 to eliminate $10,000 worth of risk. The exception would be if there are substantial other costs associated with (unknown) errors in the savings estimate.

The Value of Information. Formally, the value of improved accuracy can be calculated as the change in expected net benefits, based on the changed probabilities of various outcomes and their associated costs and benefits. For simplicity, assume that the only costs associated with an error in the savings estimate are overpayment (from the buyer's perspective) or underpayment (from the seller's perspective). Then the expected value of improved accuracy is the change in expected payments. In the example above, if the payment structure is symmetric, the expected payment is the same whether no monitoring is done, new monitoring is done with accuracy ± $15,000, or better monitoring is done with accuracy ± $7,000. In this example, if both the purchaser and the contractor believe that the savings will be $100,000 per year, and neither has asymmetric risks associated with errors in the savings, it may be reasonable to do no monitoring, and accept the initial estimate and bounds.

This case is shown in line 1 of Table 1, assuming the payment P for savings estimate $\hat{S}$ (in $1,000) is

$$P = 25 + \hat{S}/2.$$

With this formula, the payment for $100,000 savings estimate is $75,000.

Suppose, however, that the buyer believes the true savings to be $80,000/year, while the seller believes the true savings to be $120,000/year. With no monitoring, the buyer would believe the payment was $10,000 too high, and the seller would believe it was $10,000 too low (Case 2 of Table 1). Both would see value in investing some amount less than $10,000 in an unbiased measure of actual savings. However, provided the measure is unbiased, and again assuming no asymmetric risks associated with uncertainty, there would be little value to either party in investing more to tighten the bounds on the measured savings.

Now suppose that the payment structure is asymmetric. In particular, suppose that the payment is zero if the estimated savings is below $80,000, and otherwise has the same formula as before. Such an asymmetric payment structure might be adopted to provide a strong assurance that a minimum savings level would be achieved. Assume also that an error bound of (±X means that the savings estimate $\hat{S}$ has a normal distribution, with mean equal to the true savings and standard deviation X.

Case 3 of Table 1 shows the effect of this payment structure when both parties believe the true savings to be $100,000. Even though the payment at the expected savings is the same before, $75,000, the expected payment with any (unbiased) monitoring method is lower, because of the possibility that measurement error will result in a savings estimate below the threshold. If monitoring is to be done, however, the chance that the estimated savings will fall below the threshold is reduced, so that the expected payment is increased, if a more accurate measurement method is used. It is in the seller's interest, but not the buyers, to invest in more accurate measurement. However, the cost of the better measurement should be less than the expected increase in payment.

Finally, consider the case of asymmetric payment, as above, combined with disparate beliefs as to the true savings (Case 4 of Table 1). In this case, the buyer expects a decrease in the payment and the seller expects an increase if any monitoring is done. Both would have incentives to invest in monitoring to remove bias. However, the gains for either party associated with improving the monitoring accuracy are slight.

A more complete value-of-information analysis would account for other costs associated with errors in savings estimators. It would also consider a distribution of true savings, rather than a fixed assumption, and estimate the likely net benefit over this distribution.

Typically, no explicit value-of-information calculation is made in establishing monitoring and precision levels. The costs associated with errors in savings estimates are difficult to quantify rigorously. The level of investment in monitoring

Table 1. *Difference in Expected Payment by Monitoring at ± 15 or ± 7 Accuracy ($1,000)*

		Buyer's Beliefs					Seller's Beliefs				
			Expected Payment with Monitoring		Difference from No Monitoring			Expected Payment with Monitoring		Difference from No Monitoring	
Payment Structure	Case	Expected Savings Estimate	+/−15	+/−7	+/−15	+/−7	Expected Savings Estimate	+/−15	+/−7	+/−15	+/−7
Symmetric	1	100	75	75	0	0	100	75	75	0	0
$P = 25 + \hat{S}/2$	2	80	65	65	−10	−10	120	85	85	10	10
Asymmetric	3	100	69.4	74.9	−5.6	−0.1	100	69.4	74.9	−5.6	−0.1
$P = 0$ if $\hat{S} < 90$	4	80	35.5	33.9	−39.5	−41.1	120	84.8	85.0	9.8	10.0

is commonly based on rules of thumb, and a sense of what level of accuracy both parties are comfortable with. However, in cases where there are conflicts between investment guidelines and precision guidelines, or between the parties to the contract under negotiation, the value-of-information perspective can provide a useful framework for re-assessing the requirements.

In making such a reassessment, it is important to be clear about what the precision requirements mean, and how they relate to qualitative M&V objectives. Indeed, it is always valuable for both parties to understand the meaning of and motivation for the criteria.

Confidence and Precision

The above discussion of benefits and risks associated with different accuracy levels was phrased in loose terms to indicate a general principle. That principle is that investments in improved accuracy should correspond to the benefit associated with that improvement. That benefit is based on reduced expected losses or increased expected gains.

Absent from the initial discussion was reference to the confidence level associated with the indicated bounds. Specification of the accuracy of an estimate requires not only the absolute or relative bounds (± $20,000 or ± 20 percent) but also the level of confidence that the true value is within those bounds. While this requirement can seem to be a fine point, a precision statement without a confidence level defined is in fact meaningless. By allowing the confidence to be low enough, the precision bounds can be made arbitrarily tight.

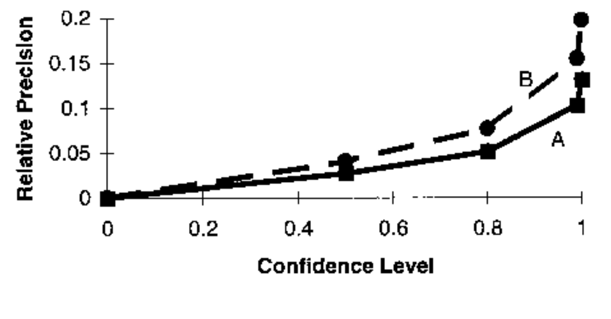
For example, suppose the precision for a particular estimate is around ± 5 percent at 80 percent confidence. Then (using the normal distribution, which is the basis for most precision calculations) the precision would be around ± 2.5 percent at 50 percent confidence, or ± 10 percent at 99 percent confidence (Line A in Figure 1). Providing the precision statement (±X) without the confidence level tells us nothing.

Likewise, comparing precision levels without knowing if they are reported at the same level of confidence is meaningless. Line B in Figure 1 is for an estimate with 50 percent worse relative precision, at any given confidence level, then that indicated by Line A. However, if the precision for A is reported at the 99 percent confidence level and the precision of B at the 80 percent confidence level, A has the wider precision: ± 10.3 percent versus ± 7.7 percent.

As discussed above, statistical precision is not the only consideration in specifying monitoring requirements. However, it is useful to understand the meaning of statistical precision measures, and the implications of different sampling strategies in terms of those measures.

Precision Standards. The need for precision standards in M&V has been the subject of some debate. In the context of evaluation, a 90/10 standard, meaning 10 percent relative

Figure 1. *Relative Precision Versus Confidence Level for Two Different Estimates*



precision at 90 percent confidence, is often invoked. This standard is included in California's Monitoring and Evaluation Protocols (California PUC 1994), and is also the basis for the sampling requirements of Pacific Gas and Electric's PowerSmart Program protocols (Pacific Gas and Electric Company 1994).

As discussed extensively by Hanser and Violette (1992) in the context of DSM program evaluation, the requirement of 10 percent precision at 90 percent confidence has been adopted in part by the extension of the Public Utilities Regulatory Policy Act (PURPA) requirements for a class load research sample (PURPA 1978). The PURPA requirement is that the sample should be designed so that the class load at any given hour is determined with 90/10 precision.

Other precision standards are applied in other disciplines. One common standard in technical publications is that results must be significantly different from zero at the 5 percent or 1 percent significance level. The requirement of an estimate different from zero at the 5 percent significance level means that a 95 percent confidence interval would not include zero. That is, for a positive estimate $\hat{y}$ the confidence interval '$\pm$' width w must be small enough that

$$\hat{y} - w > 0.$$

Therefore,

$$w / \hat{y} < 100\%.$$

Thus, significant difference from zero at the 5 percent significance level is equivalent to a 95/100 rule. Likewise, a requirement of 1 percent statistical significance level is equivalent to a 99/100 rule.

The extension of the 90/10 rule from load research to evaluation and verification has been made in several areas, but raises some questions. One question is what parameters the criterion should be applied to. A second question is the level of disaggregation at which the criterion should be imposed. In the load research context, the parameter of interest is the load at a given hour, and the level of disaggregation is the revenue class. In evaluation, monitoring, and verification, the parameter of ultimate interest may be the savings in load, energy, or energy costs at prevailing rates. The level of disaggregation is critical in the context of M&V. This level reflects—or implicitly defines—the monitoring objectives, and strongly affects the monitoring costs.

What Parameters. Measuring savings means measuring a difference in level rather than measuring the level of consumption or load itself. In general, measuring a difference with a given *relative* precision requires greater *absolute* precision, therefore a larger sample size, than measuring a level with the same relative precision. For example, suppose the average load is around 500 kW, and the anticipated savings is around 100 kW. The 90/10 criterion applied to the load would require absolute precision of 50 kW at 90 percent confidence. The 90/10 criterion applied to the savings would require absolute precision of 10 kW at the same confidence level.

In monitoring and verification, the precision criterion may be applied not only to demand or energy savings, but also to parameters that determine savings. For example, suppose savings is the product of number N of units, hours H of operation, and change C in watts:

$$S = N\ H\ C.$$

The 90/10 criterion could be applied separately to each of these parameters. However, achieving 90/10 precision for each of these parameters separately does not imply that 90/10 is achieved for the savings, which is the parameter of ultimate interest. On the other hand, if number of units and change in watts are assumed to be known without error, 90/10 precision on hours implies 90/10 precision for savings.

What Level of Disaggregation. In the M&V context, the precision standard could be imposed at various levels. The choice of level of disaggregation dramatically affects the sample size requirements and associated monitoring costs. Possible choices include

- for individual sites, where sampling is conducted within each site
- for all savings associated with a particular type of technology, across several sites for a given project, where both sites and units within sites may be sampled
- for all savings associated with a particular type of technology in a particular type of usage, across several sites for a project
- for all savings associated with all technologies and sites for a given bidder
- for all savings associated with a group of projects for a given bid program.

In general, the finer the level at which the precision criterion is imposed, the greater the data collection requirement. If the primary goal is to ensure the accuracy of savings for a project or group of projects as a whole, it is not necessary to impose the same precision requirement on each subset. In fact, a uniform relative precision target for each subset is in conflict with the goal of obtaining the best precision possible for the project as a whole.

For example, suppose that a total of 100,000 units of efficient lighting are to be installed by 10 different contractors, with different numbers of units installed by different contractors. If a precision target such as 90/10 is set for the project as a whole, totaled over all contractors, the most efficient sample will be achieved by allocating the sample to contractors roughly in proportion to the number of units installed (assuming all contractors install large numbers of units, and the distribution of savings from individual units is similar for all contractors). That is, proportional allocation will give the smallest total sample size for the targeted precision level, or the best possible precision for a given total sample size.

On the other hand, if a uniform relative precision requirement is imposed on each contractor, each contractor will be required to sample the same number of units, regardless of how many that contractor installed. The result will be a higher relative burden on the smaller contractors.

The same general principle applies in the case of a single contractor operating across several sites, or across several groups within a site. A uniform relative precision requirement for all sites or groups results in approximately the same sample size for each, regardless of its contribution to the total project savings. Optimum precision across all sites or groups together results in sample allocation in proportion to the number of units installed.

EXAMPLES OF SAMPLING REQUIREMENTS IN EXISTING PROTOCOLS

The above discussion described some general issues for setting M&V sampling and precision requirements. We turn now to some existing protocols, and consider how their sampling requirements relate to these principles. The intent is not to provide a comprehensive review of the specific protocols, but to illustrate how the general concepts can be applied, and to indicate how a lack of clarity regarding these concepts can lead to potential problems.

The NEMVP document is probably the most comprehensive and extensively reviewed set of protocols currently available. However, these protocols do not address sampling issues in any detail; nor do they suggest a formal basis for establishing sampling requirements. The discussion of accuracy there is generally broad and non-technical.

Pacific Gas and Electric's (1994) PowerSaving Partners Monitoring and Verification Procedures Manual and the New Jersey Protocols (NUBRC 1993) both include explicit sampling requirements. We use these two sets of protocols to illustrate M&V sampling issues in practice.

PowerSaving Partners Sampling Formula

The PowerSaving Partners (PSP) protocols have similar sampling requirements for lighting efficiency measures, lighting controls, and constant-load motors. The focus of the sampling procedures is on determining hours of operation.

Sampling is required within each usage area. A usage area is to be defined so that areas grouped together have similar operating hours. For lighting, areas within a group must have similar proportions of lights on within each costing period. Motors grouped together are required to have ''identical operating characteristics and/or expected operating hours. As discussed, the level of disaggregation at which a precision standard is imposed implicitly defines the monitoring objectives, and directly determines the sample size requirements and associated monitoring costs. For the PSP protocols, the level of disaggregation is the usage area.

The sample size for each usage group is given by the equation:

$$n = (zcv(y) / p)^2 \qquad (1)$$

where

n	is the required sample size
z	is the standard normal deviate for the given confidence level, specified as 1.645 for 90 percent confidence
$cv(y)$	is the coefficient of variation of hours of operation
p	is the required relative precision, specified as 0.10 (i.e., 10 percent).

This is the standard formula for the sample size required to produce a symmetric 90 percent confidence interval with 10 percent relative precision. (For ease of discussion, we ignore the finite population correction factor, which reduces the sample size for small populations.)

Implications of the PSP Sample Size Formula. For a fixed 90/10 precision standard, the only factor that will affect the sample size formula is the coefficient of variation *cv(y)*. This factor is the standard deviation of hours of operation across the different units in the usage area, expressed as a fraction of the mean hours for these units. Thus, the *cv*(*y*) is a measure of how homogenous the usage area is. The more homogeneous the group, the smaller the *cv*(*y*), and the smaller the required sample size *n*.

The PSP protocols specify 0.5 as the default initial assumption for the *cv*(*y*), in all situations, unless alternate information is presented. However, the assumed *cv* will be revisited after the first year, presumably by calculating the standard deviation of hours for each usage group from the monitoring data collected. The protocols also specify that sampling should be within groups of similar usage, and that the best possible information should be used to determine sample sizes so that the need for future modifications will be minimized. Thus, while the default assumption may give a comfortable sample size for the first year, it is important to recognize the implications for future years of the actual variability. The question of population variability, or *cv*, is at the heart of the definition of usage groups.

Suppose, for example, that there are two large usage groups A and B of equal size, each with $cv(y) = 0.5$, as indicated in Table 2.

The last line of the table shows the mean, standard deviation, and $cv(y)$ that will be found if groups A and B are combined into a single group. If the two groups are each sampled separately, the protocols (Equation 1) would require 68 observations from each, or a total of 136. However, if the two are combined into a single large group, the sample size required for the single combined group would be 237. In the first year, the default assumption would allow a sample size of only 68 for the single group (if the grouping were allowed, given the lack of similarity of expected hours). In subsequent years, however, a much larger sample size would be required based on the likely observed variations within the group during the first year of monitoring.

The reason the sample size is so much larger is that the combined group has a much larger $cv(y)$ than either group alone. The reduction of variance, and therefore of sample size required, when a large group is subdivided into smaller, more homogeneous groups, is the basic reason for stratified sampling (that is, for separate sampling within subgroups). However, the usage areas defined by the PSP protocols are not sampling strata in the conventional sense, as discussed further below.

It is likely that the initial assumption of $cv(y) = 0.5$ will turn out to be far from reality for many usage groups. A challenge for effective implementation of the PSP protocols is to establish principles for specifying realistic values of the $cv(y)$ before any sampling has been done. That is, guidelines are needed for establishing (1) usage groups within which hours are ‘‘similar;’’ and (2) when a proposed group is ‘‘similar’’ to one that has been monitored previously, to use the prior information on $cv(y)$.

The assumption of uniform *cv* across groups means that a group with smaller expected operating hours has a smaller standard deviation *s*, as indicated in Table 2. If, instead, the standard deviation turns out to be the same for two groups the *cv* will be greater, therefore the required sample size will be larger, for the group with smaller expected operating hours. For example, if the standard deviation is 1,000 hours for both Groups A and B, the *cv* and sample size for A are as shown in Table 1; the *cv* for B is now 0.17, and the sample size required only 8. This leads to the uncomfortable requirement that substantially more resources be invested in verifying the hours for the group that contributes a smaller amount to energy savings than for the one that contributes three times as much per unit. Other such imbalances can occur in other situations, so long as the protocols require the same precision for each group, regardless of the group’s contribution to total savings.

PSE&G’s Sampling Plan

The New Jersey Protocols include a sampling plan for Public Service Electric and Gas Company (PSE&G). The primary content of this plan is a set of tables indicating the sample size required for different ‘‘end-use groupings.’’ Each end-use groupings is defined by a range of on-peak and off-peak weekly operating hours. Several details are unclear from the text, including

- whether the indicated sample sizes are for selection of units to be monitored within each monitored facility, for selection of facilities to monitor within a group of facilities, or for selection of units across all facilities monitored

Table 2. *Sample Size Required for Separate and Combined Groups*

Group	Expected Operating Hours *y*	Standard Deviation of Operating Hours	Coefficient of Variation $cv(y)$	Sample Size Required *n*
A	6,000	3,000	0.5	68
B	2,000	1,000	0.5	68
A + B	4,000	3,742	0.935	237

- whether entire facilities or portions of facilities are to be assigned to groups
- what the unit of observation to be sampled is—switches, lamps, or areas
- what the source of operating hours information is for assigning facilities or portions of facilities to groups.

Defining Usage Groups

Both the PSP protocols and the New Jersey protocols specify sample size requirements for each usage group. However, both leave ambiguity as to how a usage group is to be defined. This ambiguity leaves considerable room for interpretation, and dispute, as to the total level of monitoring required for a project.

The typical sample design situation is one where a population parameter is to be estimated either to meet a specified precision level or within a specified budget constraint. In this situation, the population is divided (stratified) into homogeneous groups (strata) so that the precision level can be met at the lowest cost, or so that the highest possible precision is achieved within the budget constraint. That is, the goal is to estimate the parameter for the population; the population is divided into groups so as to achieve this goal as efficiently as possible. In this conventional sampling situation, a smaller total sample will typically be required if a multi-modal group is divided into two groups each of which is more homogeneous (has smaller variance) than the original larger group.

The usage groups in the PSP and PSE&G M&V protocols do not correspond to sampling strata in this conventional sense. The reason is that the usage groups are not defined so as to provide an efficient estimate for the population as a whole. Rather, the same precision target must be met *within* each usage group. The somewhat perverse result is that breaking the population into more usage groups will tend to increase, rather than decrease, the sample size required. Put another way, there is no statistical basis for determining an optimal or preferred definition of usage groups to meet the precision criterion, because the criterion must be met for each usage group defined. Because the protocols are vague as to how usage groups are to be defined, the ultimate precision objective, which is stated in terms of usage groups, is itself vague. That is, we have a procedure to follow once usage groups are defined, but we have no overarching principle or criterion to guide that definition.

Revising Group Assignments after Sampling

Usage areas are defined based on expected operating hours. In some cases, the monitoring data will indicate that the hours are very different from the initial assumption. In such cases, the question arises whether the area should be reassigned either during the monitoring period or prior to computing estimates. The answer is no. The sampling and estimation procedures are based on the initial groupings, however those groupings are determined. The intent of the groupings is to create homogeneous groupings, within which variances will be lower than across the total population of all affected fixtures. It is neither required nor anticipated that the initial information will always be correct. The estimation procedures will be correct as long as the sampling procedures are followed and the initial groupings are retained.

LEVERAGED SAMPLE DESIGNS

Both the PSP and PSE&G sample size requirements are based on formulas that apply when the sample mean is to be used directly as the estimate of interest. More efficient sample designs (i.e., smaller sample sizes for the same precision level, or better precision for a given sample size) are possible with a ‘‘leveraged’’ approach. Leveraging means incorporating additional information and utilizing the relationship between the measured variable and other variables known for the whole group. Leveraged approaches include regression and ratio estimators.

Regression Estimation Approaches

The PSP protocols allow the Partner to reduce the sample size by a factor of $1-R^2$ in cases where there is an established correlation R between the measurement of interest y and the initial estimate x of hours of use. That is, R is an a priori estimate of the correlation that will be calculated from a regression of the sample values of y on the corresponding values of x.

This sample size adjustment would be appropriate only if the relationship between x and y were intended to be used to improve the estimate of the population mean μ. That is, the population mean would be estimated by fitting a regression of y on x for the sample data, then using the fitted model to calculate the value of y at the (observable) population mean of x. Because this estimator utilizes additional information, the variance of the estimate is reduced. That is, for a given sample size, the regression estimator will have lower variance than the simple mean (if there truly is a correlation). Conversely, the regression estimator can achieve a fixed variance target with a smaller sample.

The PSP protocols do not suggest that such a regression estimator should be used for the sampling situations discussed, or even that one would be allowed. If the information about x will not be used in estimating the population mean hours μ, the mere existence of the relationship between x and y does not justify reducing the sample size. The protocols should not allow the R^2 correction to the sample size unless the relationship between x and y will be used in estimating μ.

If the intent is indeed to utilize the information about x in estimating y, the regression used as the basis for the a priori estimate of R^2 must be similar to the sampling situation in several ways. The protocols indicate that the value of R^2 used in the sample size calculation must come from previous similar work. The similarities must include not only the technology type and building or space type but also

- the basis for the initial estimates x of operating hours
- the installation procedures
- the range and general distribution of observed values of x and y.

Sample Design to Support Ratio Estimation

A common sampling approach for metering studies is a design to support ratio estimation. The basics of this approach are described in detail by EPRI (1983). The method is also described in most statistical sampling texts.

The idea of ratio estimation is to take advantage of information available for the whole population, rather than relying on the sample alone. The sample is used to determine a relationship between the metered observations and the other information known for everyone. This relationship is then applied to the total of the known information. Because more information is used in the analysis, a higher precision level is achieved with the same sample size, or a smaller sample is required to achieve a given precision level.

Basic Form of the Ratio Estimator

In its simplest form, the ratio estimation approach could work as follows. Let

x_i = the predicted hours of use for a circuit of lights i, determined during the installation audit
y_i = the metered hours of use for that circuit of lights
D_i = the change in watts for that circuit of lights.

Thus, $S_i = D_i y_i$ is the savings for the circuit of lights, based on the metering, and $\tilde{S}_i = D_i x_i$ is the corresponding estimate based on the initial estimate of hours of use. Further, let M denote the circuits i that were metered, and T denote the total set of all the installed lights. Then the savings S_T for the whole group would be estimated from the metering sample by the ratio estimator S_{TR}, derived as follows.

$$\begin{aligned} r &= \sum_{i\epsilon M} D_i y_i \,/ \sum_{i\epsilon M} D_i x_i \\ &= \sum_{i\epsilon M} S_i \,/ \sum_{i\epsilon M} \tilde{S}_i \\ S_{TR} &= r \sum_{i\epsilon T} D_i x_i \\ &= r \sum_{i\epsilon T} \tilde{S}_i \end{aligned}$$

Sample Size Formula Using the Ratio Estimator

Ignoring the finite population correction, the sample size required for the ratio estimator to achieve 90/10 precision can be calculated as

$$n = [(z / p) cv_s]^2 (SDR / SD)^2$$

where

z = 1.645
p = 0.1
SD = ordinary standard deviation of savings S_i across the population
cv_S = cv of savings S_i

$$SDR = \left[\sum_{i=1}^{n} (S_i - r\tilde{S}_i)^2 / (n - 1) \right]^{1/2}.$$

This sample size formula is the same as the sample size formula used for the simple mean estimator in the PSP protocols, with two changes. First, we are targeting 90/10 precision for savings, rather than for hours. Therefore we base the sample size on the cv of savings rather than on the cv of hours. This is equivalent to targeting 90/10 precision for a weighted average of hours, with weights given by the demand savings D_i.

The ratio approach could be applied directly to (unweighted) hours, rather than to savings, with completely analogous estimation and sample size formulas. Because the ultimate point of the monitoring is to assure savings, not hours, sampling and estimation based on the energy savings estimated (i.e., hours weighted by demand savings) is more appropriate.

The second change from the protocol sample size formula is that the basic formula $(z\, cv/p)^2$ is multiplied by the factor $(SDR/SD)^2$. The ordinary standard deviation SD measures the variability of the different terms S_i around their mean. The ratio deviation SDR measures the variability of the different S_i around their corresponding ratio predictions $r\tilde{S}_i$. If there is any useful relationship between the initial prediction of hours x_i and the actual metered hours y_i, the ratio deviation *SDR* will be less than the ordinary standard deviation *SD*, so that the factor *SDR/SD* will be less than one. As a result, the sample size required for the ratio estimator will be less than that required for the mean-per-unit estimator assumed for the protocols.

Ratio-Based Sampling as a Variant of the (1-R^2) Adjustment

This reduction in sample size is based on the same principle as the reduction by the factor $(1-R^2)$ allowed by the PSP protocols, except that we are relying on a ratio line rather than on a simple linear regression line. That is, the factor $(SDR/SD)^2$ plays the same role, and has essentially the same basis, as the factor $(1-R^2)$ in the protocols. Just as the $(1-R^2)$ adjustment requires an estimate of R^2 from previous work, so does the ratio-based sample size formula require an initial estimate of *SDR*. This estimate, and the corresponding sample size requirement, would be revised based on the first year of monitoring.

If the savings-weighted ratio described above is used, the expected savings S_i may vary widely across circuits. In these cases, a stratified ratio estimator using two or three size strata is likely to be more efficient. That is, the stratified ratio estimator would give better precision for a fixed total sample size, or require a smaller total sample size to achieve a given precision for the project as a whole. The idea of the stratified ratio estimator is that more of the sample is allocated to circuits with higher expected savings. The calculation of the ratio and its precision are modified to reflect this sample allocation. The stratification does not require that 90/10 precision be met within each stratum. Rather, the stratification and allocation are designed to provide 90/10 (or any desired criterion) precision for the project as a whole.

CONCLUSIONS

The foregoing discussion has identified a number of principles for establishing sampling procedures for monitoring and verification.

(1) There are both statistical and nonstatistical reasons for preferring certain sampling approaches or coverage rates. Ideally, these reasons should be articulated and prioritized.

(2) Statistical precision criteria are not meaningful without a clear statement of

- the confidence level associated with a given precision requirement
- the parameters the requirement applies to
- the level of disaggregation at which the requirement applies.

(3) Sampling and precision requirements should be based on clear objectives. In particular, the relative importance of accuracy for a project as a whole versus accuracy for individual components should be articulated, and should be reflected in the sampling requirements. The value-of-information framework, which identifies the benefits of reduced uncertainty, can be useful in defining the appropriate level of investment in monitoring.

(4) Considerable savings in monitoring costs may be achieved by applying advanced sampling procedures, including stratified sampling, ratio estimation, regression-based adjustments, and combinations of all three. However, specifying such approaches as part of a standard protocol remains a challenge; there are also costs associated with developing custom sample designs for specific situations.

The overriding principle in setting sampling requirements is that these requirements be based on clear rationale, and be unambiguously specified. The goal is to provide a level of confidence in the results and a level of cost to obtain those results that meets the needs of all parties.

REFERENCES

California PUC 1994. *Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management Programs.* California Public Utilities Commission Decision 93-05-063.

EPRI 1983. *Model-Based Statistical Sampling for Electric Utility Load Research,* EPRI EA-3286 Electric Power Research Institute, Palo Alto, CA.

Fels, M.F., K. Kissock, M. Marean, and C. Reynolds 1995. *PRISM (Advanced Version 1.0) User's Guide.* Princeton University Center for Energy and Environmental Studies, Princeton, NJ.

Hanser, P. and D. Violette 1992. ''DSM Program Evaluation Precision: What Can You Expect? What Do You Want?.'' *Proceedings of the NARUC-DOE 4th National IRP Conference,* Burlington, VT, September 1992.

NAESCO 1993. *NAESCO Standard for Measurement of Energy Savings for Electric Utility Demand Side Management (DSM) Projects, Revision 1.3* National Association of Energy Service Companies, Washington, DC.

NEMVP Technical Subcommittee 1995. *National Energy Measurement and Verification Protocol, Revision 10.* U.S. Department of Energy Washington, DC.

NJBRC 1993. *Measurement Protocol for Commercial, Industrial and Residential Facilities.* New Jersey Board of Regulatory Commissioners.

Pacific Gas and Electric Company 1994. *PowerSaving Partners: Measurement & Verification Procedures Manual.* Pacific Gas and Electric Company, San Francisco, CA.

PURPA 1978. Public Utilities Regulatory Policy Act of 1978. P.L. 95-617.