How to Hit Several Targets at Once: Impact Evaluation Sample Design for Multiple Variables

Craig Williamson, EnerNOC Utility Solutions
Robert Kasman, Pacific Gas and Electric Company

ABSTRACT

Many energy efficiency (EE) programs save energy in multiple ways, including kWh, kW and Therms. Evaluation samples are usually designed based on a single variable, with target precisions set for estimating kWh savings. This can create a problem.

While target precision levels are usually attained for kWh savings, kW and Therms savings are relegated to a "we'll get whatever precision we get" status. When a program has significant savings in more than one variable, these methods may provide only partial information and non-actionable results, undermining the fundamental purpose of evaluation. With the potential for additional types of savings and impacts, including GHG emissions (tons CO2), water saved (gallons), and other pollutant or resources, this issue will become more and more critical.

This paper will provide the background of the problem, present a recent example encountered in the industry, clarify what caused the difficulties, and then explain how careful planning and thorough sample design up front can address this problem. We'll walk through an illustrative hypothetical example that shows how to use multidimensional stratification and subset the population to correct this problem. By addressing the problem up front during sample design, improved results can be achieved with little or no additional cost. It is always cheaper to plan for things up front than to try and accommodate after data have been collected.

Overview of Current Sampling Practices

Many evaluation projects involve using a sample to estimate the achieved or evaluated savings (also referred to as the *ex post* savings) for a program. It is normally cost prohibitive to visit and audit every participant in a program, and statistical sampling allows evaluators to estimate the savings for relatively few sites, and then expand that estimate to the population of participants and also calculate the precision or accuracy of the estimated savings. It is important, and good statistical practice, to include the precision of the estimate in the reporting of the savings, to enable readers to interpret the savings more appropriately.

The first step in the sampling process is to analyze the population being studied and design the sample. This is challenging, because assumptions must be made about the eventual results in order to determine the most efficient sample and to predict the precision the sample will eventually achieve once the actual data are collected. The goal of the sample design is to reduce the variance in the savings to enable more accurate estimates from a smaller sample size. The primary method of reducing variance is to stratify the population into subgroups that have lower variance within the strata than in the overall population. In EE evaluation, the most common stratification variable is the reported savings (also referred to as the *ex ante* savings). Since the evaluated savings should be correlated with the reported savings (assuming that the initial estimates are at least reasonable), this will group customers with similar savings together.

If this correlation is not present, then virtually any evaluation method in common use today will result in poor precision. Unfortunately, evaluators can't know what the correlation is before completing the evaluation, so it is impossible to plan for this in advance, but experience and judgment can inform the design.

Another way that the precision of a sample estimate can be improved is through the use of a ratio estimate. With a ratio estimate, the relationship (correlation) between the reported savings and the evaluated savings can be leveraged to improve the precision directly. Instead of estimating the evaluated savings by itself, the sample is used to estimate the ratio of the evaluated savings to the reported savings. This ratio is the realization rate. Applying the ratio to the total program reported savings gives a more precise estimate of the evaluated savings for the whole program that just directly estimating those savings from the sample evaluated savings. It is important to note that the reported savings do not need to be accurate to improve the precision – they just need to be statistically correlated with the evaluated savings. In fact, hypothetically, if the reported savings were all about twice what the evaluated savings were, the precision of the estimate of the savings for the whole program would be excellent, because the correlation would be high. To be clear, there are two important pieces to the estimate of savings – what the point estimate of the savings is (including how close it is to the ex ante savings) and how precise that estimate is. The correlation affects the precision, but the actual ex post savings for each site determines the point estimate of the savings.

Using these two methods for improving precision, stratification and ratio estimation, a relatively small sample can be used to estimate the total evaluated savings for a program to a reasonable precision. It is common industry practice to design the sample to achieve what is known as "90/10" precision, which means 10% relative error with 90% confidence. In layman's terms, this means that you are 90% certain that the estimate (the evaluated savings, in this case) is within 10% of the true (but unknown) value. There are two components to this – the level of certainty (90%) and the size of the bounds around the true value (10% of the estimate).

However, as mentioned above, assumptions need to be made about the data that will be collected. Because we can't know up front what our estimate of the savings will be, the variance of that estimate (within each stratum), or the correlation between the reported savings and the evaluated savings, we cannot predict the precision resulting from any given sample size or stratification scheme. However, we do know the mean and variance of the reported savings within each stratum. We can use the means and variances of the reported savings with each stratum, and design the sample to achieve the target precision for estimating the reported savings without using a ratio estimate. Because the ratio estimate will always provide better precision (assuming at least some correlation between the reported savings and the evaluated savings), if we design for 90/10 (or any target precision) based on the direct estimate of reported savings, we will achieve at least that level of precision using the ratio estimate in nearly all circumstances. This approach of designing the sample based on the direct estimate of reported savings is common industry practice for estimating sample size.

This approach for sample design is very likely to result in a sample that achieves the target precision when there is a single quantity being estimated, such as kWh savings. The stratification and the ratio estimation can be based on reported kWh savings, and everything works as described above. However, nearly all program evaluations now also report peak demand (kW) savings as well as kWh savings. Stratifying based on reported kWh savings does not reduce the variance in kW savings nearly as much as it reduces the variance in kWh savings. In addition, many programs also have natural gas savings in Therms at some of the participant

sites. This causes even more problems, since there can be very little correlation between Therm savings and kWh savings, meaning that the stratification based on kWh savings does not reduce the variance in Therm savings. Not only is the correlation lower, but there are often many sites that don't have any Therm savings, and so when a sample is selected based on kWh savings stratification, there may be only a small handful of sites with non-zero gas savings.

Unfortunately, the common industry practice is to design samples to achieve the target precision for the most important quantity (usually kWh), and basically ignore the Therm and kW savings until the end, simply hoping that the precision is okay. It rarely is. This usually results in good precision for kWh, but very large confidence intervals (i.e. +/- 40%) for Therm savings. One alternative is to use the kWh savings stratification and calculate the sample size required to achieve the target precision for the other quantities (Therm and kW savings). This can work for kW savings without inflating sample sizes too much, since kW and kWh tend to be correlated, but it rarely works for Therm savings. Unfortunately, this usually results in huge and unrealistic (from a budget perspective) sample sizes because of the lack of correlation. So many fall back on the "we get what we get" for the precision of Therm savings estimates, no matter how unreasonable the resulting precision is.

In addition to sampling for multiple types of savings, there are other areas that suffer from this phenomenon of using a simple kWh sample design for more than is really possible. One such example is from an evaluation study that targeted results by measure type as well as program (Global Energy Partners, 2011), but the specifics of that approach is beyond the scope of this paper.

Recent Examples, Sampling Challenges for Multiple Savings Types

The literature includes many examples where evaluated programs had savings in multiple variables, often kWh, kW, and therms. We explore two examples here.

The first example is in the evaluation of the 2006-2008 PG&E Large Commercial incentive program (ADM Associates, *et al*, 2010). The program claimed net savings of 58,352,671kWh, 9,684kW, and 287,995Therms in the three year period. The report's sample design description states that "The design variable used in developing the sampling plan was ex-ante gross kWh savings," and the goal was that "total kWh savings could be estimated at the 90% confidence level, with 10% precision being the target" (page 4-2). The sampling plan makes no mention of kW or Therms sampling strategy, nor mentions kW or Therms parameter estimation targets. However an attempt was made to use the sample designed for kWh to also estimate kW. Not surprisingly, the results were very poor precision for kW and Therms savings. This example is described in more detail below.

A second example, Evaluation of the 2004-2005 Savings By Design Program (RLW Analytics, 2008), shows an interesting approach to address the multiple savings sampling challenge. The approach used here converted the electric and gas savings all to MBTUs, and then stratified based on MBTUs (as well as across multiple utilities). This would remove the difficulties related to Therms and kWh, but do nothing to improve kW estimation. The results would also be more difficult to interpret, especially given the expected desire to have separate estimates for Therm and kWh savings from the program, which this would not allow.

A Better Method

There is a better way. The reason that stratification improves estimation precision is that it reduces variance by grouping customers with similar levels of savings together. So we propose multidimensional stratification. Instead of stratifying only on one variable (kWh savings), stratify on multiple variables. Conceptually, this might involve, say 3 strata in the kWh dimension (low, medium, and high kWh savings), and 3 strata in the Therm dimension (no Therm savings, low Therm savings, and high Therm savings), for a total of 9 strata. Care must be taken not to overstratify, since too many strata can become unwieldy. The idea is that the kWh stratification dimension will improve the precision of the kWh savings, and the Therm stratification will improve the precision of the Therm savings estimate.

In statistics, nothing is "free" – this is not a magical way to get better precision at no cost. This can provide better precision for Therm savings, but does so because we are stratifying on Therms. However, it will usually end up increasing the required sample size somewhat, since each dimension does not help the precision of the other quantity. But it can give Therm estimates with much more reasonable precision levels, and do so for a minor increase in sample size. Or given the same sample size, it can improve the precision of Therm savings estimates significantly, while resulting in slightly less precise kWh estimates. But it allows us to plan for the precision of both quantities, instead of focusing on one and ignoring the other, while not requiring huge sample sizes.

The requirements for stratification are simple – the stratification must subdivide the population such that each participant in the population is in exactly one stratum. No customer can be left out, and no customer can be in more than one stratum. In order to sample and calculate weights, the stratum assignment must be known for all participants. Further, when the sample is selected within each stratum, all customers in that stratum must have an equal probability of selection, in order to avoid sampling bias. As long as we follow these rules, we can creatively expand stratification to capture precise savings estimation of multiple quantities.

Example of the Problem

We now look further at the PG&E Large Commercial incentive program for program years 2006-2008 (ADM Associates, *et al*, 2010) mentioned above. We use the results as our example here, which involved a sample of 46 projects across five kWh strata to estimate the savings for a population of 358 total projects. In some cases, these projects included both gas and electric savings, and the evaluation estimated the kWh, kW, and Therm savings for the program.

This is not intended as any sort of indictment on the evaluation performed for this program – the evaluators used common industry practices, as described above. But because only a small number of the projects in the population had Therm savings, the sample ended up with only 3 cases with Therm savings.

Table 1 below shows the achieved precision for each of the three quantities.

Table 1. Precision of Gross Savings Realization Rates for PG&E 06-08 Large Commercial

Quantity	Reported Savings	Evaluated Savings	Realization Rate	90% Confidence Interval
kWh	58,352,671	46,374,538	79.5%	± 12.2%
kW	9,684	8,181	84.5%	± 68.8%
Therm	287,995	60,377	21.0%	±33.8%

Source: Table 4.7, ADM Associates, *et al*, 2010.

Note that the 90% Confidence Interval for kWh is very reasonable at 12.2%, and only slightly higher than the planned 10%. It appears that the original designed sample size was 61, so apparently (the report doesn't explain this) some of the sites were not evaluated, resulting in a lower precision. However, the precision for kW and Therms are much worse, making those estimates less useful. It is surprising that with only 3 sample points that had Therm savings in the sample, the Therm precision was not far worse.

Our hypothesis is that if this sample were designed using multidimensional stratification, the precision levels would be more reasonable, particularly for the kW and Therm savings. We now describe what the evaluators did and contrast that with how we would approach this.

The evaluators stratified the population based on kWh savings, as shown in Table 2 below.

Table 2. Sample Design for PG&E 06-08 Large Commercial using kWh Savings Stratification

Stratum	Definition	Population	Designed Sample Size	Achieved Sample Size
1	kWh savings <32,000	96	2	1
2	kWh savings between 32,001 and 78,000;	82	2	2
3	kWh savings between 78,001 and 165,000	66	2	2
4	kWh savings between 165,001and 300,000	59	2	2
5	kWh savings > 300,000 (census)	55	53	39
Total		**358**	**61**	**46**

This design achieved reasonable precision for the kWh savings estimate, with the relative precision at 90% confidence at 12.2%. However, there were only three of these sample customers that had Therm savings, and those three represented only about one third of the total reported Therm savings (the sample represented about half of the reported kWh savings). The evaluators did estimate the realization rate and total program Therm savings based on the three sites, but it was not clear how they did that. They may have treated the three as a simple random sample, since there were not enough sample points to calculate a stratified estimate. Treating a stratified sample as a random sample for estimation would introduce bias, but may have been the only option available.

If during the planning phase, the sample had been stratified using two dimensions, with both Therm savings and kWh savings used in the stratification, things might have been different. We describe the method for doing this in detail with a hypothetical example below. The five strata could have been split between those with and those without Therm savings. There would

need to be more than two sample points in each half, but given a slight increase in the total designed sample size, sufficient representation of the sites with Therm savings could have been achieved. While the precision might not have been 10%, it would have been much better than 34%, stratified estimates could have been calculated which would have been valid (unlike a simple random sample estimate based on a stratified sample), and the estimate would be based on a broader sample of sites with Therm savings.

Unfortunately we were not able to obtain the population data for this study, so we could not restratify the data and calculate what the precision could have been given a two-dimensional stratification. But we would expect to see much better precision on the Therm savings estimate. This would come at a cost of a larger sample size, but we strongly believe that it would be worth it.

Of course, hindsight is 20/20 – we are looking back at what was done, and it is always easier to see ways to improve things after the fact. There were undoubtedly constraints at the time that we are not aware of. We chose this as an example because the number of sites in the sample with Therm savings was so small, that it represents a more extreme case. But there are many examples of this issue in the industry, and evaluators seem to be willing to continue to use the same approaches, which continue to result in very poor precision for estimates of non-kWh savings.

The General Approach

The general approach to estimating multiple quantities with the same sample is to stratify in multiple dimensions, with each quantity representing a dimension. In the case of gas and electric savings, this implies stratifying on kWh savings and on Therm savings. Because we were unable to acquire the population data for the above example, we have created a hypothetical example based on our experience that might be encountered in an impact evaluation, tailored to show how the method should be implemented. We describe how we would stratify based on common industry practice, and then we modify the stratification to use our proposed approach. This paper is not intended, nor could serve as, a basic primer on sample design for evaluation, so while we describe the usual approach, we don't go in depth with many details.

We assume that, as is common industry practice, savings for all the measures installed at each sampled customer site will be estimated. This is usually done because of cost consideration – much of the expense of an on-site visit is the cost of getting there, so the incremental cost of estimating savings for additional measures at a sampled site is less than the cost of estimating savings for measures at other locations. It may also be tempting to consider designing separate samples for each measure, but because of the increased cost for more on-sites, separate samples would result in much higher costs than an integrated method, such as we now propose.

Table 3 below shows the hypothetical example of how a population of 300 participants might be stratified base on kWh savings only, using common industry practice. The weights reflect the proportion of the population that is in each stratum. Often, it is more efficient not to sample proportionally, and the weights are used adjust the sample results to reflect the population distribution across strata.

Table 3. Hypothetical kWh Savings Stratification

Stratum	Definition	Population	Weight
1	kWh savings <30,000	115	0.383
2	kWh savings between 30,000 and 100,000	82	0.273
3	kWh savings between 100,000 and 200,000	64	0.213
4	kWh savings > 200,000 (census)	39	0.130
Total		300	

If a program focuses primarily on kWh savings, as many in the US do, it is reasonable to first look at stratifying to improve the precision of kWh savings estimates, and then look at Therm savings. As in the example described above, many times these programs have a majority of participants with no Therm savings at all, but a small number with Therm savings. In this case, we would recommend stratifying first based on kWh savings, but not creating too many strata – three or four kWh strata would be appropriate. As is common practice, it is a good idea to assign the participants with the highest savings to a "census" stratum, where all customers are included in the sample. This can really improve the precision. The next step is to split the population in each of the kWh strata into those with Therm savings and those without Therm savings. In some cases, there may be a kWh stratum or two with no non-zero Therm savings customers, which is perfectly acceptable. In this situation, all the customers would be in the zero-Therm stratum. If there are any individual customers with very large Therm savings, it is appropriate to assign those to their own "census" stratum, and put them into the sample with certainty. This would result in a nine-stratum design.

Say that 44 of these participants have non-zero Therm savings. Using the approach described above, we can split each of the above strata into two strata, one with Therm savings and one without. If there were 3 customers with very high Therm savings, we could pull those three out into their own census stratum. The resulting new stratification is shown in Table 4 below.

Table 4. Hypothetical Two-Dimensional Stratification

Stratum	kWh Stratum Range	Therm Stratum Range	kWh Population	Therm Population	kWh Weight	Therm Weight
1	0 to 30,000	Zero	100	0	0.333	0
2	0 to 30,000	Nonzero	15	15	0.050	0.341
3	30,000 to 100,000	Zero	70	0	0.233	0
4	30,000 to 100,000	Nonzero	12	12	0.040	0.273
5	100,000 to 200,000	Zero	55	0	0.183	0
6	100,000 to 200,000	Nonzero	8	8	0.027	0.182
7	Over 200,000	Zero	31	0	0.103	0
8	Over 200,000	Nonzero	6	6	0.020	0.136
9	Any	Over 100,000 (census)	3	3	0.010	0.068
Total			300	44		

There are a few important things to note here. The weights, which are the proportion of the population that falls within each stratum, should be different for the estimation of kWh and Therm savings, since the populations are different. However, even though the estimation process will be different, it is important to note that we use the same sample – this is not two different samples used to estimate two things, it one sample used to estimate two different things.

If we had Therm and kWh variances and means for each of the above strata (if this were based on real data, not just a hypothetical), we could calculate the required sample sizes under both of these stratification schemes to achieve a target precision. We would expect that the sample size required to achieve the target precision for kWh savings using the first stratification (kWh only, as in Table 3) to be smaller than the sample size using the second (two-dimensional, on both kWh and Therms, as in Table 4), because we are estimating two things. But in concept, the sample for stratum 1 in the first would be split between strata 1 and 2 in the second.

However, if we used the first stratification to determine a sample size needed to estimate the Therm savings to a certain target precision, we would expect that the sample sizes would be much larger than what would be required with the second stratification, since there is little correlation between kWh savings and Therm savings. The sample size would have to account for the variability in Therm savings across customers with and without Therm savings, and reflect the fact that the sample would include some customers with Therm savings and some without. But by splitting each of the kWh savings strata into the zero and non-zero Therm savings, the second scheme will more efficiently estimate Therm savings. Further splitting the non-zero Therm savings strata into a high and low Therm savings would help even more, but we would need to be careful not to have too many strata.

Unfortunately, the second scheme would result in two different allocations of sample points to the strata, one to estimate kWh savings and the other to estimate Therm savings. There would probably be a need to increase the sample in the non-zero Therm savings strata, which would increase the overall sample size required. But the benefit would be valid estimates of Therm savings that actually achieve target precision levels. And in practice, we would expect the difference to be small.

Table 5 and Table 6 below represent what we might see (again, this is a hypothetical example) if we designed samples to achieve target precision levels based on the different stratification schemes, and to estimate different quantities (kWh and Therm savings). Table 5 uses the kWh only stratification scheme.

Table 5. Hypothetical Sample Sizes using kWh Savings Stratification

Stratum	Definition	Sample to estimate kWh savings	Sample to estimate Therm savings
1	kWh savings <30,000	6	27
2	kWh savings between 30,000 and 100,000	7	18
3	kWh savings between 100,000 and 200,000	3	15
4	kWh savings > 200,000 (census)	39	39
Total		55	99

The problem is that with each stratum containing a mix of customers with Therm Savings and without Therm savings, a much larger sample size must be selected to ensure that those with Therm savings are appropriately represented. This results in the seemingly paradoxical fact that

the sample sizes required to estimate Therms are larger than the number of customers with non-zero Therms savings in each stratum. This is because the customers would be selected without regard to the Therm savings – so enough must be selected in total to ensure that the sample will include customers with and without Therm savings. Of course, if we actually used the sample size based on estimating Therm savings, we would then have very precise estimates of kWh savings, since we enlarged the sample so much.

Common industry practice would be to use the kWh-based sample design to select the 55 customers as specified above, get good precision on the kWh estimates, and then hope for the best on Therms. Note that, for instance, only 15 of the 115 (13%) stratum 1 customers have nonzero Therm savings, so there is a very good chance that none of those 15 customers would be in the sample of 6 from stratum 1. And it is possible that the only customers in the sample with non-zero consumption could be in the census stratum, and these customers, with the largest electric savings, would not be representative of the population of customers with Therm savings.

If we use a two-dimensional stratification scheme, we would expect to see sample sizes similar to what is shown in Table 6 below.

Table 6: Hypothetical Two-Dimensional Stratification

Stratum	kWh Stratum Range	Therm Stratum Range	Sample size to estimate kWh savings	Sample size to estimate Therm savings	Sample to estimate both (maximum)
1	0 to 30,000	Zero	4	0	4
2	0 to 30,000	Nonzero	3	7	7
3	30,000 to 100,000	Zero	5	0	5
4	30,000 to 100,000	Nonzero	3	5	5
5	100,000 to 200,000	Zero	2	0	2
6	100,000 to 200,000	Nonzero	2	4	4
7	Over 200,000	Zero	31	0	31
8	Over 200,000	Nonzero	6	6	6
9	Any	Over 100,000 (census)	2	3	3
Total			58	25	67

In this hypothetical example, the sample required to estimate kWh savings only is slightly larger than the sample using only the kWh savings for stratification (58 versus 55). However, when we look at the sample needed to estimate the Therm savings, we only need sample points in those strata have non-zero Therm savings. Of course, we would need to estimate both using the same sample, so we could choose the maximum sample size for each stratum from the two sample designs, shown in the final column above, which would give at least the target precision for both quantities. In fact, we may be able to reduce the sample size in strata 1, 3, and 5, since increasing the sample size in 2, 4, and 6 would improve the kWh precision somewhat.

The key thing to note is that with a small increase in sample size, using a two-dimensional stratification has the potential to deliver much better precision in Therm savings estimates. But this must be done in the planning stage. While it is possible to post-stratify after the data have been collected, the problem is that random selection would result in very few customers with non-zero Therm savings included in the sample, so doing anything would be difficult.

Of course, this is a hypothetical example, but it represents what we would reasonably expect to see in a typical evaluation project. In order to determine how much this would help Therm savings estimates, the method would need to be applied to actual population data.

For this example, we used Therm savings and kWh savings, but the same approach would apply to stratifying in two dimensions on kW savings and kWh savings. In this case, the kW savings could be split between high and low.

Conclusion

In each of the above case studies, using something other than a traditional stratification based on kWh savings was (or would have been) more efficient. In the first case study, if the population had been stratified based on Therm savings and the sample had been designed to achieve one target precision for kWh savings and less precision for Therm savings, the resulting sample size would have been higher, but the results for Therm savings would have been much more reasonable. In the second case study, the target precision levels were achieved, and cost was reduced by including more multi-measure sites without biasing the sample.

There are certainly many other ways in which innovative sample design could improve results. The key is to consider sample design a critical part of the evaluation process, and carefully assess the goals of the evaluation project when designing the sample. It is always less costly to plan for things up front and include more in the sample design than it is to try and accommodate things after the data have been collected.

Good sample design also takes time. Unfortunately, there is often extreme pressure to get the sample designed and selected, so data collection can begin. However, this rush can be costly, if it prevents a more efficient sample design that could lower the sample size or could provide better precision.

References

ADM Associates, Inc., Innovologie LLC, Marketing Excellence, Inc., C. J. Brown Energy, P.C., David Claridge, Ph. D. 2010. *Commercial Facilities Contract Group 2006-2008 Direct Impact Evaluation, Volume 1of 3 Final Report*. http://calmac.org/publications/ComFac_Evaluation_V1_Final_Report_02-18-2010.pdf.

Global Energy Partners. 2011. *Evaluation of the 2009 Energy Conscious Blueprint Program.* http://www.ctsavesenergy.org/files/Evaluation%20of%20the%20ECB%20Program%20-%20Final%20Report%208-4-2011%20final.pdf

RLW Analytics. 2008. *An Evaluation of the 2004-2005 Savings By Design Program, October 2008 Revision.* http://calmac.org/warn_dload.asp?e=0&id=2598.