Multi-Attribute Valuation for Cost Effective Evaluation of Market Transformation and other Hard to Quantify Programs

James Woods and M. Sami Khawaja, Ph.D., quantec, llc, Portland, OR

Abstract

The classical tests of cost effectiveness, TRC, RIM, UTC, are all very dependent on any program intervention being small and having no permanent change on the functioning of markets. They also require many transformations of the data to convert various benefits and costs into a dollar value. These two attributes make the classical tests difficult to use for market transformation programs and where the benefits and costs are only known with some significant uncertainty. This paper introduces multi-attribute valuation as a tool for evaluating program cost effectiveness when: programs are more complicated than a subsidy, there is disagreement among stakeholders, and when there is uncertainty about the kinds of benefits that will accrue. This method has proven to be successful in the difficult problem of evaluating the cost effectiveness of a Continuum of Care homelessness program.

Introduction

Cost benefit evaluation is a very old topic in economics, the evaluation community, and the energy community. During the demand side management (DSM) era, cost benefit analysis followed general pattern of:

- First find out how much the program is going to cost ratepayers
- Next find out how much measures cost
- Fight about how much they cost
- Then find out how Watts are saved with the new measure
- Fight about how many Watts are saved
- Then find out how many hours the measure will run
- Then fight about the number of hours.

The final steps usually consist of madly calculating the value of the energy saved and then having an argument about whether the costs represent total costs or marginal costs. The end result is a measure of cost effectiveness, either the Total Resource Cost (TRC), Ratepayer Impact (RIM), Utility Cost Test (UCT), or the Participant Cost Test (PCT), calculated to several decimal places.

DSM program evaluation already strains the limits of credibility with this methodology. Market transformation, infrastructure, and research programs are much harder than DSM programs to evaluate, since they have many benefits that need to be measured, converted into Watts, multiplied by hours, and finally converted to dollar savings. Because it is so difficult to quantify the benefits of programs that don't fit the DSM model, and the logic models of non-DSM energy efficiency programs are much more complicated, those programs are falling by the wayside in favor of DSM programs.

In this paper, we will put some bounds of where the traditional TRC/RIM/ UCT/PCT methodology is applicable and how it fails in many common cases. Once these weaknesses are exposed, we will introduce Multi-Attribute Valuation as a way of shortcutting some of these weaknesses while at the same time allowing evaluators to make more explicit the opinions of experts, results from other evaluations, and the program logic. We will also show that multi-attribute valuation is very effective program cost effectiveness technique when:

- The program is more complicated than a simple direct subsidy, e.g., market transformation
- There is disagreement among the stakeholders about how the program functions
- There are so many uncertainties that the classical tests produce highly doubtful results

Why TRC, RIM, PCT and UCT Are Not All that They Seem

The classic tests, TRC, RIM, UCT, and PCT are all well thought out, well defined, easily understandable, intuitive, and completely unrecognizable as a welfare measure by anyone that comes from public economics. These tests do not resemble the welfare measures, consumer and producer surplus, compensating variation, or equivalent variation, which are used throughout public economics. . They don't come out of that literature.

Public economics focuses on how changes in prices, subsidies, lump-sum transfers, and other market interventions affect the efficient allocation of goods and services. They take into account how public programs are funded, in this case by increasing the price of electricity, and how this affects the decisions of the population. The classic tests assume that all transfers, the funding of the program, and the subsidies given on any measures, all appear in a lump sum and have no effect on prices.[1]

This means that the overriding assumption in all the classic tests is that programs are small, have no effect outside their immediate sphere, and produce no permanent change in markets. In other words, by applying these tests to market transformation programs, you are assuming that the programs are incapable of producing changes outside their immediate sphere, and any changes that do occur are temporary and fleeting. Clearly, if you have a market transformation program, this is not the ruler that you want to have used when your performance is measured.

These are, however, exactly the rulers that you want to use when your program is small, dealing with only a few buildings, or having an impact on only a very small part of the market. In each of these cases there will be no changes in any prices and no large changes in behavior; the assumptions in the tests are exactly in line with the assumptions of your program.

These small programs, where the tests are appropriate, expose another weakness of the classical tests; they don't adequately capture the effects of uncertainty that is more evident the smaller programs become. Think about the variables that go into cost effectiveness estimates of a simple DSM program giving away a free CFL. You need to know:

- The cost of the CFL
- The lifetime of the CFL
- The hours of operation and the wattage difference
- Administration costs
- Avoided costs of the energy saved

Each one of these variables is estimated; they are random. Consider the simplest parameter, the lifetime of a CFL. Yes, you may say that, on average, a CFL lasts seven years. What you are really saying is that half the CFLs will last between two and ten years.[2] That is a big range, and the range gets larger for longer-lived measures. When a measure has a 15-year life, half the measures will last between 1½ and 21 years. This range gets even wider when you are uncertain about the expected measure life.

Its true that the classical tests are usually conducted with some kind of sensitivity analysis, changing one variable at a time, but even that does not capture the full uncertainty of these values.

[1] Standard Practice Manual: Economic Analysis of Demand-Side Management Programs (1987)

[2] This is a result of using the exponential distribution to model the lifetime of a measure.

Clearly this range is too small to account for measure life sensitivity, not to mention that the estimates could be wrong for more than one variable. Uncertainty, or the variance in the parameters, does not act like regular numbers. When you subtract one uncertain number from another, you do not subtract the uncertainty – the uncertainties are added. Uncertainty grows with every parameter that is added to a calculation.

This provides yet another reason not to use the classical tests with market transformation programs – they have many more parameters. Because they have many more parameters, the uncertainty about their costs and benefits is even larger, even when we know all the individual parameters with relative certainty.

We can never eliminate the uncertainty, nor should we. But the degree of uncertainty makes it difficult to produce *one* cost effectiveness assessment. That one cost-effectiveness assessment must satisfy all the stakeholders, even when there is disagreement about how a program will work and there are many different ways that the program can produce energy savings and welfare improvements. The more parameters you have, the more disagreement there will be about those parameters. Market transformation programs generate parameters rather quickly.

To reiterate, because market transformation programs are more complicated than simple DSM programs and are intended to make permanent changes in markets and behavior, they should not be evaluated with the classical tests because:

- The classical tests assume there is no *permanent* change and that all changes are small
- The uncertainty in the classical tests expands rapidly with the number of parameters and market transformation programs have many more parameters than simple DSM programs.
- Stakeholders are likely to argue about the assumptions and evaluation of market transformation programs. These arguments are less likely to be resolved because *one* evaluation must be produced.

Multi-Attribute Valuation

Multi-Attribute valuation is a way of sidestepping many of these issues. Some evaluators may have already been exposed to multi-attribute valuation, perhaps without being aware. If you have seen a Request for Proposals that contained a score sheet, where so many points are gained for clarity of the proposal or the quality of the work plan, you have seen a multi-attribute valuation schema. That is generally the way that multi-attribute valuation works – many dimensions are combined into a single, composite valuation.

The Core of Multi-Attribute Valuation is Elicitation

Returning to the example of proposal evaluation schema, most of these are created by an individual sitting down and deciding on the weights that should be applied to each measure to create the weighted average, but the theory is much more flexible. There is no need for a weighting function – the attributes do not need to be added or multiplied together. Furthermore, there are many ways to construct weights other than having someone generating indices. The more commonly accepted method is to create hypothetical examples and then either[3]:

- Allow decision makers to compare one example to another and decide which is better
- Ask the decision makers to provide an over all evaluation of the hypothetical program

[3] These techniques are illustrated in many places. Some good examples are, Triantaphyllou, Evangelos, Sanchez, Alfonso (1997), Teasley, C E III (1994), and Daniels, Richard L. (1992)

- Allow the decision makers to change a parameter to meet a certain objective, such as making two hypothetical cases equally valuable

Once these data are collected, various statistical procedures can be used to tease out the criteria that were used to make these decisions. The general name for all these procedure is "preference elicitation."

In energy efficiency program evaluation, we can apply these tools in several ways. The first step, no matter how you proceed afterwards, is to create a list of program attributes that:

- Are cheap to measure
- Have an unambiguous interpretation
- Are reasonable indicators of a "good program"
- Reflect all parts of the program logic model

In order for the preference elicitation exercise to be maximally effective, the number of indices has to be as limited as possible yet provide a reasonable indication of all the other services that are not explicitly used as an index.

In an energy efficiency program, these attributes may take the form of a Watt reduction, but they can also include, for example, architect inquiries on a phone line, order placement time, or any other measure of success. From this initial list, we can make decisions about what data will be collected, based on cost, and which should not be collected due to collection difficulty.

The second step is to elicit the way that these attributes relate to a "good" or cost-effective program. We have defined two ways of doing this. The first method requires several other programs that are similar but not exactly the same as the program being evaluated. The intent of this method is to use cheap and easy-to-measure criteria as a precise indicator of what a holistic program evaluation would yield. The second method polls a panel of experts with hypothetical examples to find out what constitutes a "good" or cost-effective program – we elicit a standard.

Method One: Using other Program Evaluation Results

In the first method, we depend on having several similar programs in existence that have already undergone a holistic evaluation with a full logic model or experts that are intimately familiar with the other programs.

Over the course of the evaluation of those programs, many measurements of various targets, intermediate indicators, and final results will have been taken. We can collect the unique indexes that were used in the evaluation. If the programs have been evaluated, we will also know if they have been cost effective or not.

A panel of experts that are familiar with the program currently being evaluated can look over the indexes that are defined in the other programs and decide which are appropriate for the evaluation of the local program, based on the general applicability of the index, the cost, uncertainty inherent in the index, and the index's ability to adequately sample the various parts of the local programs logic model. It's unlikely that all the programs will have collected exactly the same indices, but many will be similar and can be converted from one to the other. Even with this conversion, there will still be holes in the data where one program has indices 1 through 7 and another program is missing, say index 2, and yet another is missing index 6.

Normally this would be considered a fatal flaw, since observations with missing data are removed before analysis. This procedure, called list-wise deletion, can result in biased estimation with very high variances. There are, however, a group of multiple imputation techniques that enable you to impute the missing values from those that are available.[4] This body of techniques is what really enables close comparisons between similar programs that were evaluated under slightly different conditions.

[4] Rubin, D.B. (1987)

Actually using these techniques without violating some of their underlying assumptions requires the care of a good statistician and is beyond the scope of this paper.

Once the data on the other similar programs have been collected, standard discrete choice statistical techniques can determine what characteristics contribute to a "good" or cost-effective program. Then, by using the very same statistical model and the attributes collected about the program we wish to evaluate, we can produce a statement about the probability that the local program is "good" or cost effective.

The added bonus is that the statistical model will also allow us to report uncertainty about the cost effectiveness of the local program, something seldom reported in energy efficiency cost effectiveness evaluations.

Method Two: Creating Hypothetical Programs

In the second method, a panel of experts is presented with several hypothetical sets of attributes. By making decisions about which example program is better than another, adjusting one attribute so that both example programs are the same, or by some other technique, the beliefs of the experts about how the attributes contributes to "goodness" can be ascertained.

The key to this technique is in not exhausting the experts. That means keeping the number of hypothetical programs down to the smallest number necessary to get good estimates in the statistical model that will be used to evaluate the cost effectiveness of the local program.

We have tried this technique in an evaluation of a Continuum of Care homelessness program. We chose as our participants local stakeholders and national experts in Continuum of Care programs. Throughout the process, we had to continually update the indices we used to describe the hypothetical programs because we were not able to get them into the same room to get up-front agreement on the indices. As a consequence, we had to use a simple imputation technique to fill in some indices that were used by one expert but not by another.

In order to reduce the possibility of exhausting the experts, the hypothetical programs were created using an orthogonal main effect experimental design (OME). The OME design allows for efficient experiments when there is no interaction between the any indices in the model. In the case of the homelessness programs, this means that the jobs training effort was no more or less effective when there was strong or weak participation in drug and alcohol programs.

One of the issues that we discovered in the Continuum of Care cost effectiveness evaluation that probably carries over to energy efficiency programs is that the experts' ability to evaluate the hypothetical programs depended a great deal on how familiar they were with the local homeless population and the services that were provided. Those that had little familiarity with the local programs typically focused on only a few of the indices, while those with more "street knowledge" tended to consider a larger set of indices when making their cost effectiveness assessments.

Once the preference elicitation exercise is completed, several different kinds of analysis can follow. If all the observations are pooled, the preferences that are used to evaluate the "goodness" of a program represent a compromise between the stakeholders. Each of their beliefs is contained in the data and the commonalities in the way they use the indices to evaluate cost effectiveness arise with the strongest statistical relationships. Those indices that they disagree most strongly about rarely present themselves as being statistically significant predictors of cost effectiveness.

It is possible to not pool the observations and to create separate models for each of the stakeholders or each group of stakeholders. By using these separate models the program cost effectiveness can be evaluated from each stakeholders' point of view. This means that the beliefs and the assumptions of the stakeholders about which indices are important and which are not can be made very explicit. As a consequence, it is possible to look at the evaluation report and see which commissioners and which board members do not think that one particular index is important and which members are most inconsistent about their valuations.

How Multi-Attribute Valuation Sidesteps the Problems of the Classical Cost Effectiveness Tests

The main weakness of the classical tests that we initially described were that:

- The classical tests assume there is no *permanent* change and that all changes are small
- The uncertainty in the classical tests expands rapidly with the number of parameters and market transformation programs have many more parameters than simple DSM programs
- Stakeholders are likely to argue about the assumptions and evaluation of market transformation programs. These arguments are less likely to be resolved because *one* evaluation must be produced.

Multi-attribute valuation has few of these weaknesses. First, the modeling technique is very flexible, which means that there are no assumptions about the permanency of change. Second, much of the uncertainty inherent in cost effectiveness evaluation is canceled out in the final valuation. Third, stakeholders can have an evaluation tailored to their responses and preferences, and if those preferences are contradictory they can be exposed to help resolve the conflicts between stakeholders.

How Multi-Attribute Valuation Does Not Assume "No Permanent Change"

The classical tests are often criticized for being very inflexible, they are creatures of their assumptions and one of the largest assumptions is that programs are small. Multi-attribute valuation is meant to side step this assumption because it is constructed from various parts of a program logic model and logic models can depict any program, even when there are several parallel mechanisms where the energy efficiency programs can produce benefits.

Returning to our trial of this technique in the homelessness program, we had to find a link between the services that were applied and the ability of the program to reduce the incidence of homelessness, the duration of stay, and the chances of returning to homelessness. To construct the indices we went to the program logic model and pulled out some of the measurable delivery targets, e.g. bed nights, some intermediate results indicators (new jobs), and a measure of costs. The hypothetical programs were then constructed to find the tradeoff between these objectives. So, there could be circumstances were the hypothetical program had a relatively low level of service delivery, but it resulted in higher than expected intermediate result and costs were moderate. Our experts responded as to how cost effective they thought this program was in meeting the terminal goals of reducing the incidence of homelessness, the duration of stay, and the chances of returning to homelessness.

Our homelessness program example truly does not have the "no permanent change" assumption, nor is there any mention of kWh saved, free riders, or any other variables usually associated with classical cost effectiveness evaluations. If multi-attribute valuation can be used to determine the cost effectiveness homeless programs, it can be used in market transformation energy efficiency programs.

How Multi-Attribute Valuation Reduces Uncertainty Growth

Multi attribute valuation reduces uncertainty in two ways. First, it does not require chain calculations, e.g., kW to kWh to dollars saved; rather it combines indicators in one step. Secondly and much more paradoxically, multi-attribute valuation can reduce uncertainty by providing less detail. These two ways of reducing uncertainty are actually linked.

The best way of explaining the link is through the analogy of a thermometer. Think of two ways of determining the average velocity of the molecules in a body of water. You could construct a sophisticated device that measures the velocity of individual molecules. This device could then record a large sample of the individual molecule velocities and you could, taking into account the measurement error of this sophisticated device, determine the average velocity of the molecules. Alternately, you could stick a three-dollar thermometer in to the water, stir it gently, and read off the temperature.

Multi-attribute valuation acts like the thermometer. By choosing indexes that are before and after any complicated calculations you can effectively skip them and still approximate the relationship between the observable delivery targets and intermediate results indicators or any other group of variables. Any parameters that are estimated in multi-attribute valuation will reflect both the uncertainty in the model and the error from the approximation.

The key assumption is that the approximation error and the error in the indices used in the multi-attribute evaluation are significantly less than the error that is achieved by adding more parameters to the model. This is actually a classic statistical problem: the more parameters that are added to a model, the less you know about them individually. At the same time, as you add parameters, you can increase the likelihood that the statistical model will not predict well because it so closely follows the existing data.

How Multi-Attribute Valuation Can Stop Arguments among Stakeholders

One of the most arduous parts of classical cost effectiveness measures is the negotiation of the cost, savings, and measure life estimates with the various stakeholders. These are supported by facts, but the interpretation of these facts can differ from stakeholder to stakeholder. Market transformation programs compound this problem because, not only are the basic facts in dispute, but the way that the program will induce permanent changes in the marketplace may also be disputed.

As we have demonstrated before, when using multi-attribute valuation with hypothetical programs, the different participants can have different valuation responses, and the cost effectiveness calculations can be shown from the point of view of each of the participants. There is an additional advantage in creating separate cost effectiveness models for each of the stakeholders; it can help highlight inconsistencies in an individual stakeholder's assessments and help surface assumptions about program performance that the stakeholder has not be able to communicate to the other stakeholders.

An example of this may be that one stakeholder does not understand that providing a subsidy to one measure will actually increase the relative cost of a similar efficient measure. The implication is that some of the benefits are taken back through the reduced purchases in the unsubsidized measure. In this case, if the stakeholder's assessments of the hypothetical programs reveals no relationship between the effects in the other measures market and cost effectiveness, a discussion about that potential relationship can bring the stakeholders to better agreement about the effectiveness of the program.

In a similar case, one of the stakeholders may have in mind a different mechanism for the success of the program's market intervention. Once the multi-attribute analysis reveals this difference, the program logic held by that one stakeholder can be explained to the other stakeholders. Once again, the stakeholders are more likely to be in agreement when the unvoiced assumptions are surfaced and discussed.

Other Benefits

Besides correcting the shortcomings of the classical test, multi-attribute valuation has other potential befits that can be exploited. If the multi-attribute valuation is implemented by using the full logic model evaluations of other programs, those evaluation results and the opinions of the experts that participated in and conducted those evaluations can be harvested to inform the local evaluation.

This means that, if there is a large enough body of results from similar programs, multi-attribute valuation can be used as a low-cost substitute for a full logic model evaluation of a program. This can result is significant cost savings, since a full logic model evaluation typically costs two to three times that of a simple cost effectiveness evaluation.

Summary

Multi-attribute cost effectiveness evaluation can be implemented for energy efficiency programs and is particularly useful in situations where:

- The program is more complicated than a simple direct subsidy, i.e., market transformation.
- There is disagreement among the stakeholders about how the program functions.
- Where there are so many uncertainties that the classical tests to produce highly doubtful results.

It also avoids some of the shortcomings of the classical tests of cost effectiveness, i.e., TRC, RIM, UCT, and the PCT. The weaknesses in the classical tests are particularly evident when evaluating market transformation programs because:

- The classical tests assume there is no *permanent* change and that all changes are small
- The uncertainty in the classical tests expands rapidly with the number of parameters and market transformation programs have many more parameters than simple DSM programs.
- Stakeholders are likely to argue about the assumptions and evaluation of market transformation programs. These arguments are less likely to be resolved because *one* evaluation must be produced.

The first step in the process of using multi-attribute valuation for cost effectiveness is to create a series of indices that:

- Are cheap to measure
- Have an unambiguous interpretation
- Are reasonable indicators of a "good program"
- Reflect all parts of the program logic model

Then by either collecting data from other similar programs that have full logic model evaluations, or by constructing hypothetical programs and having experts evaluate them for cost effectiveness, a model that defines the contribution of these attributes to the cost effectiveness of the program can be defined.

This model can then be used to evaluate the cost effectiveness of the current program and provide uncertainty bound on the evaluation, another service that is not provided by the classical tests.

Bibliography

Standard Practice Manual: Economic Analysis of Demand-Side Management Programs (1987) California Public Utilities Commission.

Triantaphyllou, Evangelos, Sanchez, Alfonso (1997) A sensitivity analysis approach for some deterministic multi-criteria decision-making methods, *Decision Sciences*, Winter 1997, 28(1):151-194

Teasley, C E III (1994) Bridge over troubled waters: The limits of judgment in decision making, *Public Productivity & Management Review*, Summer 1994, 17(4):325-334

Daniels, Richard L. (1992) Analytical Evaluation of Multi-Criteria Heuristics, *Management Science*, Apr 1992, 38(4):501-513

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.