State-of-the-art in behaviour change program evaluation

Kathleen Gaffney
DNV GL
Palace House
3 Cathedral Street
SE1 9DE London
UK
Kathleen.gaffney@dnvgl.com

Agapi Papadamou
DNV GL
Palace House
3 Cathedral Street
SE1 9DE London
UK
agapi.papadamou@dnvgl.com

Ken Agnew
DNV GL
122 West Washington Avenue, Suite 1000
53703-2715 Madison, WI
USA
ken.agnew@dnvgl.com

Keywords
behavioural change, feedback, impact evaluation

Abstract
This paper provides an overview of evaluation methods being used to evaluate a new generation of opt-in behaviour programs in the US. Specifically, this paper provides the following:

- An overview of the issues and challenges of evaluating opt-in behaviour programs,
- A preliminary look at a Californian program as a concrete example of this kind of program, and
- Recommendations relating to the evaluation of opt-in behaviour change programs.

For purposes of this paper, a behaviour program is one that attempts to influence customers to change their physical assets (energy-related investment behaviour) and/or their operations and dwelling use behaviour through information and encouragement methods, without directly providing financial assistance. These programs include audit-only programs, targeted information programs, and comparative information programs. Behaviour change programs may, however, include encouragement to participate in other programs that do include incentives and assistance.

Recent opt-in behavioural change programs using randomized controlled treatment (RCT) assignment have provided a model for unbiased evaluation based on differences between "participant" and "non-participant" consumption. However, most program designs are not easily compatible with random assignment, and require alternative evaluation methods. While audit and information programs have existed for decades, evaluation of these programs using advanced consumption data analysis methods is still in its early days. Such evaluation approaches are the most promising for comprehensive evaluation. At the same time, much work remains to assess the effectiveness of various techniques to quantify and mitigate self-selection effects.

Introduction
The purpose of this paper is to recommend approaches for evaluation of the new generation of opt-in behaviour change programs. To that end, the paper provides the following:

- An overview of the issues and challenges of evaluating opt-in behaviour programs.
- A preliminary look at Pacific Gas & Electric's (PG&E) Progressive Energy Audit Tool (PEAT) as a concrete example of opt-in behaviour change programs.
- Recommendations relating to the evaluation of opt-in behaviour programs.

The initial overarching goal of this analysis was the evaluation of the PG&E's PEAT program. The PEAT invites customers to log in to a web portal that collects information about the household and household energy consumption characteristics and uses this information to support the participant in saving energy. These tools provide advice ranging from simple behaviour changes, to low-cost changes to make to one's house or business, to suggestions of other utility rebate programs that support limited or comprehensive retrofits at the household or business. Goals for this kind of program include generating low cost savings and developing a richer utility-customer interface.

PEAT is an example of the new generation of opt-in behaviour programs that are being rolled out in California and other US states.

This paper does not provide a full-blown evaluation of the PEAT program. At the time of planning the evaluation there was insufficient data and budget to provide a full evaluation. Instead the ultimate goal was to establish how to best evaluate the PEAT program. Opt-in behaviour programs offer a particular challenge to the evaluator, and the evaluation community is just coming to terms with this challenge.

Consumption data analysis and behaviour programs

Recent behavioural programs using randomized controlled treatment (RCT) assignment have provided a model of unbiased evaluation based on differences between "participant" and "nonparticipant" consumption. However, most program designs are not easily compatible with random assignment, and require alternative evaluation methods.

All evaluations that cannot use a true RCT design are dependent on quasi-experimental methods, or even non-experimental methods. In these cases, potential bias in the construction of the counterfactual is always an issue that needs to be acknowledged and at least qualitatively assessed. This potential for bias exists for any evaluation method, including self-reports, choice modelling, and consumption data analysis. The potential is of particular concern in contexts where the program effect of interest is relatively small. In these situations, the uncertainty related to potential bias can be as large as the estimate of interest. This is a concern for most opt-in behavioural programs.

Distinguishing true program-related savings from a pre-post consumption analysis is not a simple process. There are a variety of factors that can be confounded with program effects in this estimate including weather, economic trends, system shocks, and the general summation of the remaining site-specific, non-program, pre-post changes discussed above. Various methods have allowed us to control for some or all of these effects to a degree that has made billing analysis an accepted evaluation methodology for programs including whole-building retrofit, low-income, and efficient heating, ventilating, and air conditioning (HVAC) programs.

A primary reason the potential biases are accepted for this kind of evaluations is that the magnitude of expected savings ranges from 10 % to above 20 % of consumption. Bias of up to plus or minus 1 to 2 percentage points could be ignored given the much greater magnitude of the savings. Furthermore, if bias was explained by an inability to fully control for a general upward trend in consumption then savings estimates would be safely underestimated. On occasion, unexpected occurrences such as Hurricane Katrina or the stock market crash of 2008 made it difficult to produce a reasonable savings estimates, but that was the cost of an otherwise reliable evaluation method.

Behaviour programs require a further level of scrutiny of the challenges of consumption data analysis. The most basic reason for this is the relative magnitude of expected behaviour program savings. The potential bias becomes a much greater concern when the savings are relatively small – less than 5 %. As the potential but un-measureable bias increases as a percentage of savings, the validity of that savings estimate is undermined.

The need for added scrutiny goes beyond this. Opt-in rebate programs typically generate savings through the installation of measures. While human behaviour may affect the exact level of savings generated by a measure, the behaviour-related variation is likely to be small compared with the consistent, measure-based average savings. A focus on tracked installed measures simplifies the billing analysis. The shift from pre- to post-program is a shift from one mechanical steady-state to another. This simplifies the characterization of consumption in both periods. Compared to this, the post period consumption change of a behaviour program is much more complicated. Change occurs over time, may not be consistently maintained, and may be relatively modest. It is much more difficult to distinguish a variable program effect from exogenous trends and weather variability.

In particular, opt-in behaviour programs necessitate more scrutiny on the process of self-selection into a program and the implication for estimated net effects. Self-selection is not a new issue or one that is confined to behaviour programs, but like the more general potential bias issues associated with billing analysis, concerns related to selection, failed to eliminate the problem. It was precisely the presence of RCT behaviour change programs that re-inserted the consideration of selection into the discussion of billing analysis.

SELF-SELECTION

The discussion of self-selection frequently takes on a degree of mystery. The concept is challenging, and can be explained from a number of angles.

Often the discussion is in purely statistical terms. For example, Imbens and Wooldridge discuss "unobserved covariates that are correlated, both with the potential outcomes and with the treatment indicator" (Imbens 2010). Alternatively, one can refer to the endogeneity of the treatment decision. These approaches focus on how self-selection, the process of decision-making by members of a group, may lead to selection bias in an attempt to statistically measure an effect related to those decisions.

This statistical context is essential to understanding how self-selection can result in biased estimates from a statistical model. However, this technical exposition is not necessary to understand the mechanics and effects of self-selection more generally. It is possible to understand how self-selection causes trouble in simpler terms.

Self-selection is a process whereby customers decide for themselves whether to participate in a program or not. It is present any time customers are opting in or out of a program. As a result, self-selection is present for almost all programs, unless a strict RCT design is followed. Randomly selecting customers and then taking opt-in or opt-out from the random selection leaves us with self-selection.

RANDOMIZED CONTROLLED TREATMENT (RCT)

The extensive discussion of the potential biases in a non-RCT experimental design provides a useful foil for an explanation of the importance of an RCT design. By randomly assigning a control group, the RCT approach explicitly maintains a population that should provide, on average, a perfect counterfactual with respect to every concern that we have discussed. With randomized assignment, there is:

- No concern about the imperfect process of matching. It is unnecessary.
- No concern that appropriate sites remain in the potential comparison group population. The populations are similar by construction.
- No concern about time varying characteristics. The populations are similar by construction.

For a web-based, opt-in program, a recruit and deny approach is an option. When a customer signs up for the program, they are informed that a randomly assigned subset of users will have their involvement postponed by a year. The utility could even provide a reward to those customers denied entry that is approximately equal to the average expected savings. This approach maintains the essential random assignment, but does so among customers that have shown interest in taking part in the program. This approach should support an unbiased estimate of program savings given an interest in the program. That is, there is internal validity within the population of interested customers. This approach is operationally challenging because of the built-in necessity of non-trivial level of customer deferral. It requires meeting cost-effectiveness requirements with a smaller pool of active participants. It also means upsetting some customers who do not want to wait. This approach is also not ideal from an external validity perspective because it is unclear if subsequent program participants will have similar saving characteristics.

Case study: PG&E progressive energy audit tool

The initial purpose of this project was to scope out an evaluation for the PG&E PEAT. The scope of work made it clear that there was insufficient budget or data with which to conduct a comprehensive evaluation of the program. At the time, however, we did envision a report that spent time describing the PEAT program and discussing the specific aspects of this program that could be integrated into an impact evaluation.

Our subsequent research changed our focus to wider horizons. The more general question of whether opt-in behaviour programs can be evaluated replaced the more specific question of how the PEAT program could be evaluated. Given the range of opt-in programs underway or envisioned in California, this wider horizon seemed justified.

Research into opt-in behaviour-program evaluation techniques demonstrated that program specific data are in many ways not relevant for the evaluation techniques in use today. The only program specific input used in any of the approaches discussed in this report is the time of opt-in.

Our exploration of the PEAT program reinforced this conclusion. The program collects substantial amounts of data from the participants. These data are primarily useful for segmenting and characterizing customers with respect to their reported demographics and audit responses. These data allow for a rich picture of program participants but are of limited use to the impact evaluation because they are limited to participants alone.

One possible use of program data for impact evaluation did surface. Presently, the date of opt-in enters into regressions as 0/1 indicator variable for all customers. Examination of program data ought to support the creation of indices that correlate with variation in engagement across participants. For instance, a customer who returns for a second time displays considerably more investment than the customer who never returns after the initial log-in. We could capture this information in the participation variables that enter into the models.

Conclusions and recommendations

Recent behavioural change programs using RCT assignment have provided a model of unbiased evaluation based on differences between "participant" and "non-participant" consumption. However, most program designs are not easily compatible with random assignment, and require alternative evaluation methods.

Any evaluation that cannot use a true RCT design is dependent on quasi-experimental methods, or even non-experimental methods. In these cases, potential bias in the construction of the counterfactual is always an issue that needs to be acknowledged and at least qualitatively assessed. This potential for bias exists for any evaluation method, including self-reports, choice modelling, and consumption data analysis. The potential is of particular concern in contexts where the program effect of interest is relatively small. In these situations, the uncertainty related to potential bias can be as large as the estimate of interest. This is a concern for most opt-in behavioural programs.

While audit and information programs have existed for decades, evaluation of these programs using advanced consumption data analysis methods is still in its early days. Such approaches are the most promising for comprehensive evaluation. At the same time, much work remains to assess the effectiveness of various techniques to quantify and mitigate self-selection effects.

Based on the review in this paper, the following methods are recommended:

- A combination of the VIA method and matched comparison group should be used, depending on the specific characteristics of the program.
- VIA can be used provided:
  - Opt-in dates are spread out over the evaluated program months.
  - Customers who opt in at different dates are similar.
  - Savings estimates for longer-term participants are supported by sufficient data.
- Site-specific weather normalization needs to be incorporated into VIA models. With the varied weather that characterizes the California service territories and the variable and trending nature of the program effect, it is not reasonable to expect monthly fixed effects to fully control for weather variability over time. The savings itself is likely to be weather-dependent and this effect needs to be captured in the model.
- Even with the above conditions met, inclusion of a matched comparison group with the VIA model should be tested as part of the analysis. It is more difficult to develop a rolling comparison group based on consumption in the immediate pre-program period but this approach should be developed and the comparison group integrated into a combined

VIA/difference-in-difference approach. With this approach, three alternative results should be reported: 1) VIA with no comparison group, 2) rolling matched comparison group, and 3) the combined VIA difference-in-difference. None of these approaches completely addresses the concern regarding biased savings estimates, but the three results will be indicative of the sensitivity of the results.

- For opt-in programs that start on a single date, a matched comparison group with weather normalization must be used without VIA. Such programs are not amenable to the VIA approach.
- Matched comparison groups should be treated sceptically if there is a substantial portion of the participant group that has few good matches among the non-participants. Signs of poor ability to match include large "distances" between participants and their matches, large differences in average consumption between participants and matched comparison, or extensive re-selection of the same non-participants as matches.
- To support the quantitative measurement of consumption effects, a qualitative analysis of program data should provide evidence of changes due to the program. For example, this could reflect subsequent visits to the site with indications of the completion of planned energy savings tasks.
- Evaluation of PG&E's PEAT program should begin with participant analysis, to assess which approach will be more appropriate. That is, examine the opt-in timing distribution and characteristics of customers joining at different times.
- Other programs' claims for "joint savings," if any, need to be subtracted from the consumption-based estimate of behaviour program savings when assembling a total portfolio claim. The joint savings are the incremental claimed savings from other programs that were induced by the behaviour program. Both programs contributed to the savings, but they can be counted only once for the portfolio, and typically they are counted by the non-behavioural program. The joint savings subtracted should be the incremental claim under the other program(s).

IMPROVING AVAILABLE EVALUATION METHODS

At the same time that the next evaluation is conducted, research should be done to improve on these methods and our understanding of what works. Two key steps are:

- First, explore improved matching algorithms based on key consumption parameters regardless of method pursued. Rather than matching on a series of monthly consumption values, it may be more effective to match on a limited set of parameters that characterize consumption patterns. Site-specific weather models produce heating and cooling change per unit temperature change, break-even temperatures for use of heating and cooling, non-weather-sensitive usage, and diagnostics indicating how stable or variable the consumption pattern is. When daily or hourly data are used, matching on a reduced set of usage parameters, including the indication of variability, may be much more effective than minimizing distance to overall load pattern.
- Second, existing RCT program data sets should be mined to better understand the extent of selection bias with particular analysis approaches. These datasets provide an unbiased estimate of savings with which to compare the various matching and modelling approaches for opt-in behaviour programs. In particular, this is a way to quantitatively measure the effectiveness of constructed comparison groups in general, and compare across comparison group methodologies more specifically.

References

Abadie, A. and Imbens, G.W. "Bias-corrected matching estimators for average treatment effects." Journal of Business & Economic Statistics 29.1 (2011): 1–11.

Richardson, V., Agnew, G. and Goldberg, M. "Evaluating Opt-In Behavior Programs: Issues, Challenges, and Recommendations" White Paper, California Public Utilities Commission – Energy Division, 2014

Imbens, G.W. and Woolridge, J.M. "Recent Developments in the Econometrics of Program Evaluation," Journal of Economic Literature 47. 2009.

Harding, M. and Hsiaw, A. "Goal Setting and Energy Conservation." Under revision, Journal of Economic Behaviour and Organization, 2013.

Opinion Dynamics. Massachusetts Three Year Cross-Cutting Behavioural Program Evaluation Integrated Report: Prepared for Massachusetts Energy Efficiency Advisory Council & Behavioural Research Team. Final. Waltham. 2012.

Provencher, B and B Glinsmann. "I can't use a Randomized Controlled Trial – NOW WHAT? Comparison of Methods for Assessing Impacts from Opt-In Behavioural Programs." International Energy Program Evaluation Conference: Getting it Done! Evaluation Today, Better Programs Tomorrow. Chicago, IL. 2013a.

Provencher, B., et al. "Some Insights on Matching Methods in Estimating Energy Savings for an Opt-In, Behavioural-Based Energy Efficiency Program". International Energy Program Evaluation Conference: Getting it Done! Evaluation Today, Better Programs Tomorrow. Chicago, IL. 2013b.

State and Local Energy Efficiency Action (SEEAction) Network. Evaluation, Measurement, and Verification (EM&V) of Residential Behaviour-Based Energy Efficiency Programs: Issues and Recommendations. Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman, Lawrence Berkeley National Laboratory. http://behavioranalytics.lbl.gov. 2012.