Impacts of Feedback Programs: Generating Comparable Impacts across Varying Program Design Models

Anne Dougherty, Opinion Dynamics, Oakland CA
Katherine Randazzo, Opinion Dynamics, La Jolla CA
Alan Elliott, Opinion Dynamics, Oakland CA

ABSTRACT

A number of feedback models have become prominent in our industry—from online portals to reward systems to "expert" advising—and most models have advanced far beyond providing paper reports. However, few third-party impact and process evaluations have been conducted on this newer breed of behavioral programs, and there is a dearth of literature and methodological guidance on how to evaluate behavioral program efforts that do not use pure experimental design.

In this paper, the authors discuss how we used quasi-experimental evaluation approaches to evaluate feedback programs, with a particular emphasis on developing a rigorous counterfactual to reduce self-selection bias.

We begin the discussion by theorizing the forms of self-selection bias present in opt-in feedback programs. We then discuss and augment counterfactual approaches promoted in current protocols, and how each approach does or does not address different forms of self-selection bias present in opt-in energy programs. To conclude, we provide three real-world program examples that have used different counterfactual approaches to estimate savings for the opt-in feedback programs cited above.

Introduction

In any impact evaluation, an evaluator's primary objective is to identify a rigorous counterfactual[1] to support estimates of the program effects (State and Local Energy Efficiency Action Network 2012, 7). This is also the most challenging task. In the simplest terms, a counterfactual is a point of comparison used to determine *what would have happened in the absence of the program.* The difference in energy use between the counterfactual and the participant group is used to estimate the net savings achieved by the program.

In the case of opt-in feedback programs, the evaluated savings are extremely sensitive to the counterfactual design due to the relatively small effect sizes (ranging from 0 to 9% savings per household). If the counterfactual contains significant bias (i.e., there are significant differences between the participant group and the counterfactual), then the savings estimates can vary dramatically and are difficult to replicate. In effect, we have little confidence in the accuracy of our results.

To address these concerns, there has been a strong movement in energy program evaluation to call for Randomized Control Trials (RCTs)[2] or other purely experimental evaluation techniques for all feedback programs. The SEE Action Network's recent protocol "Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations" strongly recommends using purely experimental methods to establish a counterfactual RED (2012, p.5). By randomizing eligible customers into treatment and control

[1] A rigorous counterfactual is a comparison group that is identical to the treatment group in both characteristics and time period. The only difference is that the counterfactual does not receive the treatment. Theoretically, the ideal counterfactual would be the exact same individual(s) or household(s) in the exact same time period that does not participate in the program (treatment).

[2] A Randomized Control (led) Trial (RCT) is an experimental program design in which treatment and control groups are randomly assigned, which results in unbiased program energy savings estimates.

groups, evaluators can feel relatively confident that they have accounted for potential observed and unobserved differences in the treatment population and its counterfactual, the control group.

This paper puts forward a number of experimental approaches that can be used to achieve this level of confidence, including the following:

- Recruit-and-delay
- Recruit-and-deny
- Random encouragement design RED (2012, p.14-15)

While we agree that these three evaluation approaches are ideal for estimating rigorous savings estimates, they are very difficult to implement in utility program settings for the following reasons:

Challenges to Program Design

1. **Program design and evaluation are not usually integrated processes.** In most cases, program implementers do not, and often cannot, integrate evaluators into their program design. To implement these methods, third-party evaluators need to design and validate the experiment to ensure that the experimental conditions are met and maintained through implementation.
2. **Experimental designs favor simple treatment approaches.** Experimental designs are difficult to implement in practice. For this reason, they favor very simple outreach tactics, such as direct mail, that can be carefully controlled in the field. As programs attempt to engage customers through other tactics, such as mass media outreach, experiments cannot be maintained and/or cannot differentiate the program effect from other market effects (in effect, mass media efforts implemented by the program become the new baseline).
3. **Mass-market approaches are necessary to meet energy and carbon goals.** Quite simply, these programs have the ability to realize significant energy savings. For this reason, they are powerful tools to generate direct savings (those attributed solely to the program) and indirect savings (those gained by channeling customers into other programs). The ability to scale these programs to mass-market models is vital to the continued success of feedback programs, but also critically important to bolstering the effectiveness of other programs. Purist evaluation approaches should not be used as an argument to scale these programs to even more savings across a population.

Challenges to Customer Engagement

1. **Experimental models can violate equity mandates.** Experimental designs purposefully withhold treatment from eligible customers in order to retain control groups. This can violate the equity mandates of publicly funded programs.
2. **Delaying or denying treatment can adversely affect customer satisfaction.** It goes without saying that offering and then denying treatment to customers is a less-than-ideal way to foster customer satisfaction. For this reason, most utilities and regulators will not approve these implementation approaches.

To remain forward-thinking, we must actively work toward developing rigorous quasi-experimental, non-random approaches to evaluate opt-in and mass-market feedback programs.[3] To do this, evaluators need to come together to produce more careful and skillful counterfactual approaches to address selection bias. This is particularly challenging because quasi-experimental

[3] A quasi-experiment is an empirical approached that is used to estimate the causal impact of an intervention on participants. Quasi-experimental research designs have the same goals as randomized control trials, but do not utilize random assignment to create a counterfactual.

research designs require careful consideration of the program design, the target population, and the hypothesized causal influences and effects of the program treatment. In this way, it is as much of an art as it is a science, and it requires embedding evaluation planning into program design and implementation.

In this paper, we discuss four quasi-experimental approaches used in feedback program evaluation. We discuss the trade-offs of each method to demonstrate that each method merits careful consideration. *In practice, we have found that there is not an obvious hierarchy to the methods themselves; rather, there is an ideal counterfactual approach based on the program design, implementation, and data availability.* For this reason, we argue that evaluators and regulators should not default to a single method or approach based on assigned levels of rigor. Rather, evaluators should seek to obtain the highest level of rigor possible through careful use of evaluation approaches.

To illustrate this point, we begin by hypothesizing the selection biases present in opt-in populations. We then discuss the methods proposed in existing protocols, and suggest possibilities to augment these methods. For each method, we discuss the pros and cons with respect to managing selection bias. We provide real-world examples of how counterfactuals have been successfully developed under varying program models and conditions to illustrate the careful attention required to develop a defensible quasi-experiment. The program examples include: (1) an opt-in energy information display program; (2) an opt-in report-based program; and (3) an opt-in online feedback portal using smart meter data. We conclude with a call for evaluation frameworks, rather than protocols, greater collaboration and regulatory support for integrating evaluation and program design, and more active efforts to build a literature base for quasi-experimental study designs for feedback programs.

Forms of Selection Bias in Feedback Programs

Identifying and controlling for selection bias in energy programs may be the single most vexing challenge faced by evaluators. For opt-in programs, self-selection bias is probably the most difficult to identify and control. Self-selection bias is introduced when customers are allowed to select to enroll themselves into a participant group, in contrast to opt-out models where customers are assigned participation. This results in a non-probability sample of treatment customers. To account for this bias, a comparison group is drawn to closely replicate the treatment population and the factors that may have led to self-selection, both observable (such as energy use, region, and demographics) and unobservable (such as propensity to participate and internal drive to save energy). The important differences are those that would affect energy usage and energy-use responses to historical/economic events. These differences must be identified and controlled for in the comparison group to generate a defensible study where these types of pre-existing differences between the program population and the comparison group do not lead to inaccurate results or conclusions. Below, we discuss both observable and unobservable influences on selection bias.

Observable Influences

Observable influences are those characteristics that are easily measured and quantifiable. When considering selection bias, there are a number of observable variables that are often used to draw comparison groups. Here, we break them into participant and time-variant characteristics.

- Participant characteristics: Those characteristics that define an individual or program participant that may have an effect on self-selection and energy use patterns.
    - Demographics, such as age, income, gender, and education
    - Baseline/pre-period energy use, such as average annual usage, seasonal usage, peak demand, seasonal peak, and energy intensity
    - Geographic region, which reflects and co-varies with a number of influences on participation, such as climate, culture, and demographics

    - Housing stock, such as age and size of home, building envelop, heating fuel, etc.
    - Tenancy, such as rent or own
- Past program participation in either energy or non-energy programs
- Time-variant characteristics: those characteristics known to vary over time.
    - Weather
    - Other seasonal effects, such as program campaigns and seasonal consumption trends (e.g., back-to-school, winter holidays, etc.)
    - Economy, changes in overall consumer confidence, spending, job growth, unemployment, etc.

Unobservable Influences

Unobservable influences are those characteristics that must be collected through primary data and/or may be hypothesized to co-vary with observable characteristics. When accounting for selection bias, unobservable influences are the most difficult to account for in comparison group selection. These also can be broken into participant and time-variant characteristics.

- Participant characteristics
    - Pre-existing savings trajectories, indicators that one is on a "path to save," and that the program may not have prompted participation. In the case of feedback programs, this effect can indicate that customers are using the feedback tool to track or monitor existing behaviors or trajectories, rather than to prompt new action. In addition, potential comparison group members may match participants on usage history because they have already started where the corresponding participants have not.
    - Propensity to opt-in, one's intrinsic inclination to engage in program initiatives
- Time-variant characteristics
    - Changes within household during evaluation period, such as the loss of a family member, changes in occupancy, vacations and holidays, etc.
    - Other market influences, such as social, political, or cultural trends or backlashes, changes in product availability, etc.

In an ideal world, we would have sufficient time and funding to control for many, if not all, of these factors by applying multiple methods. However, evaluators are most often working under limited timeframes and budgets, and must make careful trade-offs to best account for the biases hypothesized to influence participants. For this reason, it is critical to understand the finer details of the program model and theory, how customers were targeted, how it was implemented, and what effects have been observed to date before selecting an approach.

Quasi-Experimental Methods Bias-Control Assessment

Below, we discuss five methods that have been used to develop comparison groups for feedback programs, and the pros and cons of each in accounting for the aforementioned biases: (1) Matching on Observables; (2) Propensity Scoring Matching; (3) Future Participants; (4) Variation in Adoption; and (5) Inverse Mills Ratio and Instrumental Variable Approaches. Our discussion is not a statistical one, but a conceptual one based on statistical issues and solutions across methods.

1. Matching on Observables

The Matching on Observables method constructs a comparison group based on matching observable (or possibly unobservable) characteristics RED (2012, p.17). Of all comparison group approaches, this is the most commonly employed approach. The matching process can include one or

multiple variables used to develop either a one-to-one match of participants to comparison group members, or an overall participant to comparison group balance. At minimum, Matching on Observables requires pulling matched non-participants on observable variables from a uniform time period prior to program participation.

Pros of this method in controlling for selection bias. Of all approaches, Matching on Observables is the most accessible and cost-effective comparison group approach. It can be leveraged for most program models and designs, and with the proliferation of purchasable secondary data at the household level, evaluators have the opportunity to develop more sophisticated, multi-variable matches. These multi-variable approaches improve the likelihood that the comparison group is controlling for key unobservable factors, such as one's propensity to opt-in.

Further, this method can be used in situations where early and late participants are different from one another or where savings trends occur prior to participation. In both cases, careful use of the Matching on Observables method can be used to find comparison group members to serve as match for these effects.

One of the most compelling approaches within the approach of matching is to match by usage history, not only by average consumption over a year, but also by matching month-to-month usage over the course of one or two years. Thus, participants and non-participants appear to have reacted similarly to economic and historical events during that period, at least in terms of their energy usage.

Cons of this method in controlling for selection bias. This method uses a pool of non-participants from which to find "matches" to participants. The primary problem with this approach is that evaluators have no way of knowing whether non-participants, who look similar to participants on observable characteristics, are actually similar in terms of unobservable characteristics such as propensity to participate in general or in energy-savings programs.

2. Propensity Scoring Matching

The Propensity Scoring Matching (PSM) method constructs a comparison group by selecting non-participants based on characteristics that best predict participation (Rosenbaum & Rubin 1983). This is a statistical matching approach that uses logistic regression or other methods to identify the variables that best predict participation. In so doing, this method aims to identify the variables that are most predictive of participation to develop a more robust matching process.

Pros of this method in controlling for selection bias. Similar to the Matching on Observables process, this is a cost-effective method to developing matches, and it carries with it the same pros of standard matching approaches with additional benefits. Through predictive statistical models, this approach enhances the confidence that one is selecting comparison group members on variables that are the most meaningful. As a result, researchers have greater confidence that they have identified and matched on variables that predict participation, thus potentially better accounting for self-selection bias.

Cons of this method in controlling for selection bias. Unlike the Matching on Observables approach, this method requires multiple variables and large sample sizes to run propensity models. Often, evaluators have a limited number of variables for analysis and do not have the budget to purchase or collect additional data needed to run logistic regressions.

In addition, our experience has shown that this approach can *over-select* customers with the greatest propensity to participate. By selecting only on the variables that are most predictive of participation, this method does not adequately replicate or represent the natural variation in the participant population across these variables. As a result, it produces a comparison group that is

skewed on these variables when compared to the participant population. For this reason, the comparison group can be unbalanced. For example, 20% of participants may have high usage, while only 5% of non-participants do. A comparison group using this difference may produce a comparison group with 30% who have high usage, thus making the comparison group even more non-comparable than it would have been with random selection. We further detail our experience with this approach in the following examples section.

3. Future Participants

For programs that have cohort-based participation models, Future Participants can be used as a comparison group pool. This approach uses later cohorts of participants as the comparison group for earlier cohorts, typically comparing them across longer periods in time. This approach can be used for rolling enrollment as well as punctuated or temporally concentrated periods of time

Pros of this method in controlling for selection bias. The primary benefit of this approach is that it accounts for the propensity to opt-in because we know that Future Participants have a similar inclination to participate in programs. In addition, this approach allows for long periods of matching across multiple seasons, which can account for seasonal effects or triggers that might have prompted participants to opt-in.

Cons of this method in controlling for selection bias. This method on its own may not be sufficient for a comparison group. For example, Future Participants should be examined against observable characteristics, first and foremost energy use, to determine whether earlier cohorts of participants are inherently different than later cohorts of participants. This is because the approach assumes there are no inherent differences between early and late adopters, and also may not account for differences in programmatic efforts year-over-year, such as targeting or recruitment strategies that would affect who opts-in to the program.

4. Variation in Adoption

Similar to the Future Participants approach, Variation in Adoption (VIA) is a quasi-experimental approach that takes advantage of the pre-participation billing periods of later adopters as a rolling comparison group. This approach has been used in other evaluation efforts and has been applied in an energy context (Harding & Hsiaw 2011; Hoynes & Schanzenbach 2009; Lovenheim 2009; Opinion Dynamics Corporation 2012; Reber 2005). The later adopters serve as points of comparison until they enter the program as participants, and then are dropped from further analysis in the control group. The primary difference between this method and the Future Participants method is that the comparisons occur on a rolling basis, and the matching between current and future participants is done month-to-month rather than across an entire year.

Pros of this method in controlling for selection bias. Similar to the Future Participants method, this approach accounts for the propensity to opt-in. Further, one can argue that the closer matching process, month-to-month versus year-to-year, can better control for the differences between early and late adopters.

Cons of this method in controlling for selection bias. This method has similar constraints as the Future Participants approach, with a few additional maintained hypotheses that are difficult to meet in real-world settings. First, the model assumes that there are no fundamental differences between early and late adopters, which is often untrue, and has been proven on more than one occasion to not hold true when comparing pre-period usage patterns among potentially matched participants in evaluations (Opinion Dynamics Corporation 2013). For this reason, the method also requires

relatively uniform and consistent opt-in patterns over time, which does not account for differences in program efforts and their effects, such as targeting or recruitment strategies that can affect who, how, and when participants opt-in to the program. Finally, the model also assumes that there are no pre-period savings trajectories among participants, and that all savings effects begin with treatment RED (2012, p.5).

5. Inverse Mills Ratio or Instrumental Variable Approaches

The Inverse Mills Ratio (IMR) and Instrumental Variable (IV) approaches use multiple variables to explain the differences between a participant and a non-participant group, usually identified through another process (such as Matching on Observables or Future Participants). The approach then subjects the probabilities from that model to an algorithm that produces the Inverse Mills Ratio (separately for participants and non-participants), which is then added to the billing regression analysis to control for the unobservable factors that differentiate the participants, with the intention of controlling for self-selection bias (Heckman 1979). The variables used for the IMR and IV approaches can be produced from secondary data or primary data collection efforts, such as survey research.

Pros of the IMR/IV method in controlling for selection bias. The primary advantage of IMR is that it controls for many unobservable variables of self-selection bias by using observable variables that are collected from primary or secondary data, as well as survey research.

Cons of IMR/IV method in controlling for selection bias. Unlike other methods, this method requires that researchers identify the right variables to use in the adjustment algorithm. For the IV approach, these variables must be demonstrated to be correlated with participation but cannot be strongly correlated with the participation effects, namely energy savings. For this reason, identifying the right instrument or variables for either approach can be very challenging and expensive. Finally, this approach cannot be used in a fixed-effects, cross-sectional time series model since the IMR/IV does not vary over time. Thus, other factors that were not measured are left uncontrolled.

Three Examples of Applied Counterfactual Selection for Feedback Programs

Here, we provide three real-world examples of how counterfactuals have been successfully developed under varying program models and conditions to illustrate the careful attention required to develop a defensible quasi-experiment. The program examples include:

1. An opt-in energy information display program, using a multi-staged Matching on Observables approach
2. An opt-in report-based program, first attempting a PSM approach and ending in a Matching on Observables approach
3. An opt-in online feedback portal using smart meter data using the VIA approach

Matching on Observables – A Multi-Staged Approach

The Opinion Dynamics Team conducted an evaluation of a residential opt-in energy information display (EID) program. To conduct our impact analysis, we utilized the Matching on Observables method for developing the comparison group. This method was chosen for a number of reason, namely: (1) participants entered the program at roughly the same time period, thus eliminating the option of the VIA method (which we describe below); (2) the evaluation was due before the next cohort of participants was solicited (eliminating the option of using Future

Participants); and (3) the evaluation budget and scope would not allow for the exploration of a IMR or IV approach.

We began our matched comparison group selection by first matching on monthly pre-period usage using a matching approach. This approach generated the 10 closest comparison group customer matches for every participant based on monthly pre-period usage.

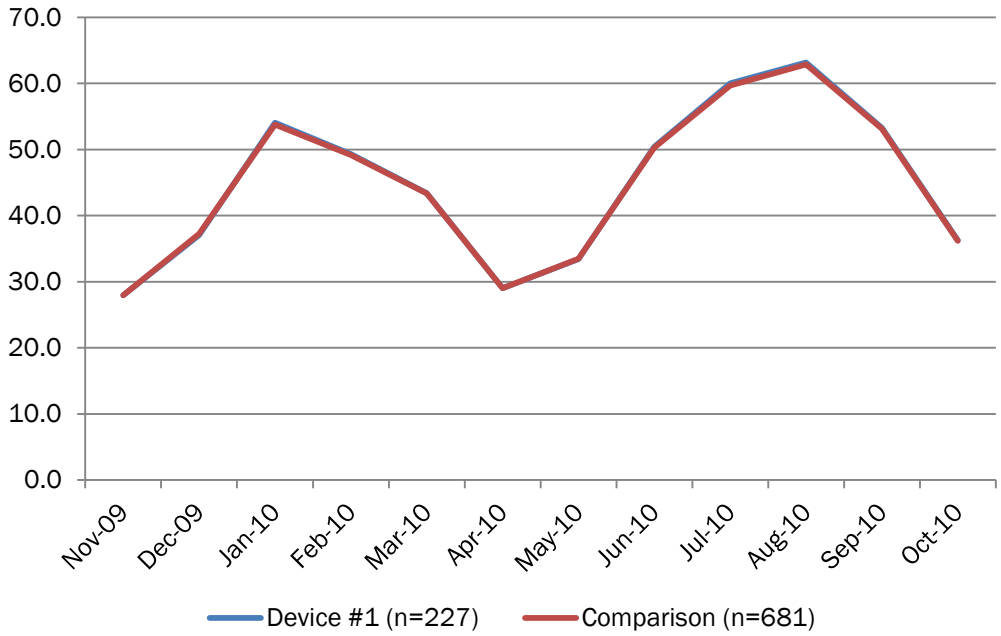

Figure 1. Monthly Pre-Period Average Daily Consumption Matches between Participant and Comparison Group Customers[4]

To further refine our matching process, our team purchased household-level demographic data through a secondary data source. This data was used to narrow our comparison group to three comparison group customers for every one participant. As demonstrated in the table below, this added step allowed our team to refine our matches on key variables known to affect customer engagement in utility programs, namely:

- Education, which skewed higher among participants than selected comparison group members
- Age, where comparison group members were significantly more likely to be over the age of 75
- Length of residency, where comparison group members were significantly more likely to have been in their home for over 20 years

These additional adjustments on demographic data enhanced our confidence in the comparison group, and this final 1:3 household participant to comparison group method was used as the counterfactual for the impact evaluation.

[4] Note the chart does include both a blue and red line, however the matching process created an exact match which makes the lines virtually indistinguishable.

Table 1. Comparison of Participant and Comparison Group Members on Demographic Data before and after Demographic Matching

Matching on Usage Only			Matching on Usage and Demographics		
Education			**Education**		
Less than high school	19%	22%	Less than high school	19%	19%
High School	19%	21%	High School	19%	17%
Bachelor's Degree	60%	54%	Bachelor's Degree	60%	62%
Master's Degree	2%	2%	Master's Degree	2%	1%
Age			**Age**		
19-24	1%	0%	19-24	1%	0%
25-34	10%	9%	25-34	10%	9%
35-44	17%	17%	35-44	17%	20%
45-54	28%	25%	45-54	28%	26%
55-64	28%	23%	55-64	28%	27%
65-74	13%	16%	65-74	13%	13%
75+	3%	11%	75+	3%	5%
Length of Residence			**Length of Residence**		
0-1	1%	1%	0-1	1%	1%
2	2%	6%	2	2%	5%
3	4%	6%	3	4%	6%
4	7%	5%	4	7%	5%
5	8%	7%	5	8%	7%
6	9%	5%	6	9%	7%
7	8%	4%	7	8%	5%
8	2%	4%	8	2%	3%
9	9%	3%	9	9%	4%
10	5%	4%	10	5%	5%
11-15	18%	15%	11-15	18%	20%
16-19	9%	10%	16-19	9%	9%
20+	19%	29%	20+	19%	21%

Propensity Score Matching (PSM) Test

Opinion Dynamics conducted a test of the PSM method to determine its effectiveness in generating a comparison group for an opt-in energy report program. In this program, interested customers conduct a brief survey of their home's characteristics and then begin receiving monthly reports benchmarking their usage against similar homes and their own usage in the previous year. The reports also provide tips and suggestions for home improvements.

As noted earlier, our research has shown that this method, while promoted as a reliable approach, can over-select for variables that are shown to predict action. Our attempts to apply this approach demonstrated similar results, as we found a significant skew among PSM comparison

group participants on key variables known to effect participation. These variables included baseline usage, age, income, and education.

We then compared the results of this approach to those generated using a multi-staged matched comparison group selected by the program implementation team. The goal here was to determine how the PSM method compared to a multi-staged matching on observable process. The results are show in the following charts.

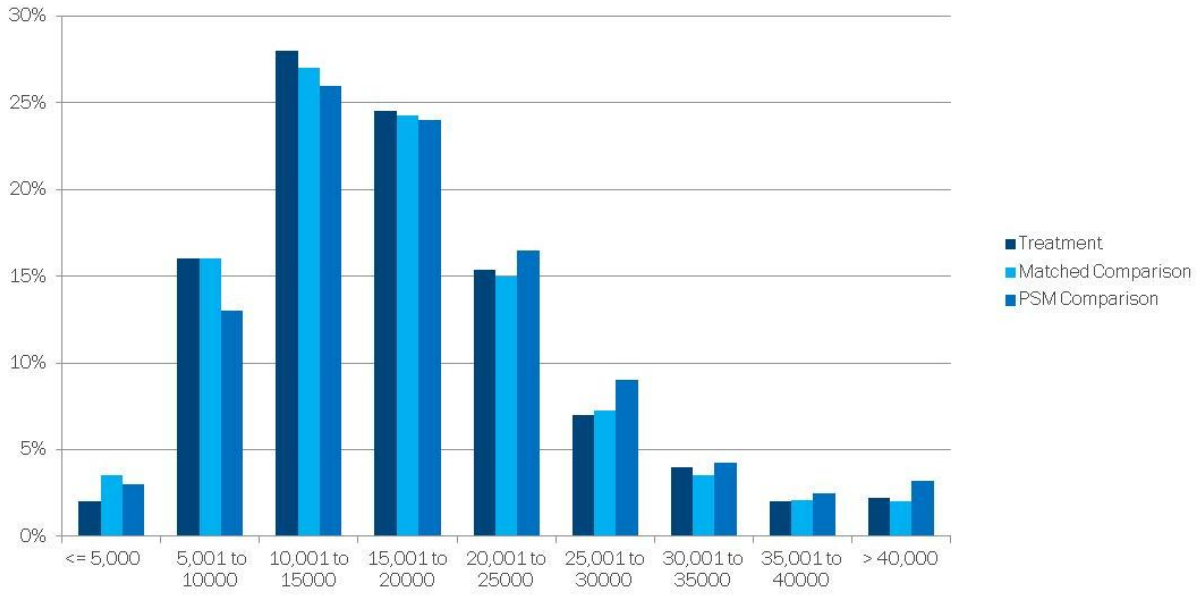

Figure 2. Annual Pre-Period Usage Comparing Participants to PSM and Matching on Observables Comparison Groups

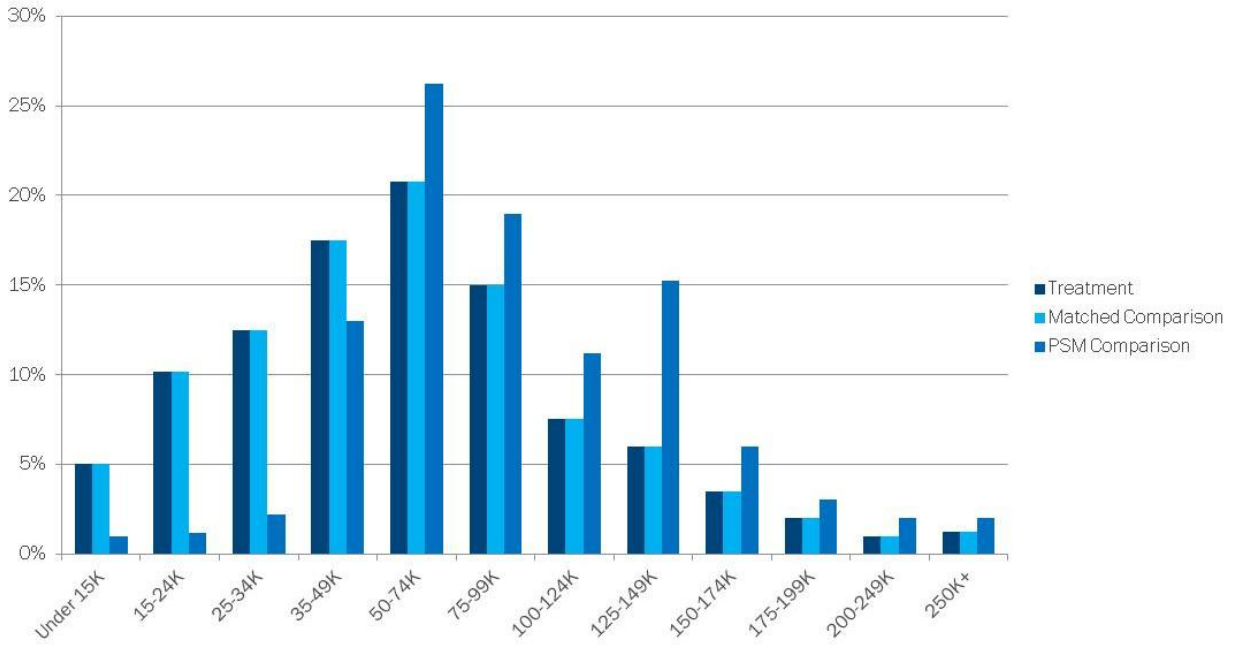

Figure 3. Annual Income Comparing Participants to PSM and Matching on Observables Comparison Groups

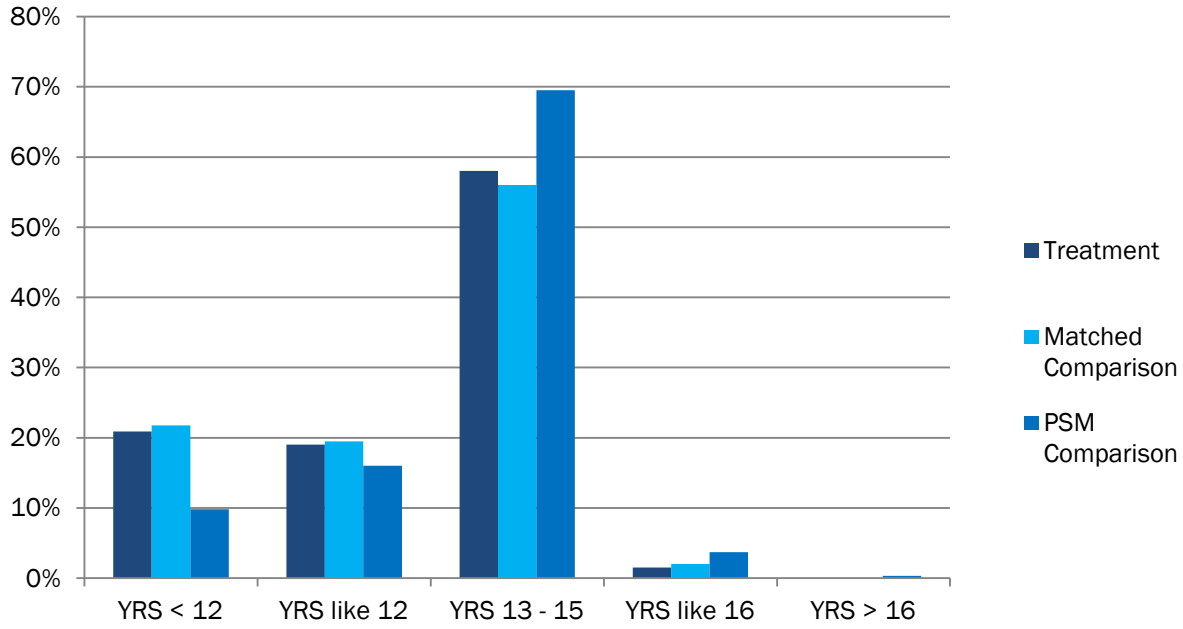

Figure 4. Years of Education Comparing Participants to PSM and Matching on Observables Comparison Groups

Based on this analysis, the team chose to use the Matching on Observables approach due to better alignment of the treatment and comparison groups. By doing so, we ensured that we were not biasing the comparison group *against* the program by over-representing customers with a high propensity to opt-in in the comparison group.

Variation in Adoption (VIA)

The Opinion Dynamics Team used the VIA approach to estimate savings for an online energy feedback program, MYMETER, implemented by Accelerated Innovations for Beltrami Electric Cooperative. This program is offered market-wide to customers at a Midwestern utility. Customers, upon sign up, receive feedback on their energy use over time and against select benchmarks. Customers can also examine the resulting savings from particular actions taken in the home using energy benchmark functions.

We generated savings for this program using a two-way VIA fixed-effects model. This model controls for the fixed effects of household characteristics as well as the month of observation. The savings estimates per household were then later explored by usage quartiles to examine whether savings effects differenced by baseline usage. The table below provides our results.

Table 3. Percent and Total Change in kWh Usage Overall and by Pre-Participation Usage Quartile

Statistic	Baseline Usage Quartiles				
	1, Low	2	3	4, High	All
Number of Participants	327	327	327	326	1,307
Total Daily KWh Change for Group	958	186	-713	-2,396	-1,965
Lower Bound of 90% Confidence Interval	512	-224	-1,135	-2,965	-3,617
Upper Bound of 90% Confidence Interval	1,404	596	-292	-1,827	-314
Average kWh Change per Household	2.93	0.57	-2.18	-7.35	-1.50
Percent Change	16.1%	1.6%	-4.0%	-8.1%	-3.1%

Note: Negative numbers indicate savings; positive numbers indicate increases in usage.

The savings found here using the VIA method were accepted by the local regulators as validated savings attributable to the program.

Conclusions

Quasi-experimental methods are necessary to support emerging and growing feedback program approaches. However, there are few studies that directly outline and convey the trade-offs between different counterfactual approaches for feedback programs for quasi-experimental methods. Further, all methods have pros and cons, and there is not true hierarchy in approach. Also, each has limitations in terms of budget and model requirements, and multiple methods may be needed to develop a defensible counterfactual.

As an industry, it is important to advance this conversation with an eye toward developing frameworks for decision-making in evaluation. Rather than protocols, such frameworks account for the differences in program design, implementation, and data available to evaluators under real-world conditions. In addition, current protocols should be expanded to reflect the trade-offs implicit in applied evaluation practices. We need to consider moving to a rigor-based framework that reflects program design differences.

To move forward, it is also important to reconsider how evaluation engages with program design and implementation. The finer details of a program approach, as well as the implementation activities, are critical to selecting a defensible counterfactual. For this reason, we need to consider fewer restrictions in using evaluation in the design phase of programs.

References

Harding, M., and A. Hsiaw. 2011. “Goal Setting and Energy Efficiency.” Working paper.

Heckman, J. 1979. “Sample Selection Bias as a Specification Error.” *Econometrica* 47:153–161.

Hoynes, H., and D. Schanzenbach. 2009. “Consumption Responses to In-Kind Transfers: Evidence from the Introduction of the Food Stamp Program.” *American Economic Journal* 1:109–139.

Lovenheim, M. 2009. “The Effects of Teachers’ Unions on Education Production: Evidence from Union Election Certifications in Three Midwestern States.” *Journal of Labor Economics* 27:525–587.

Opinion Dynamics Corporation 2012. MA Three-Year Cross-Cutting Integrated Behavioral Program Evaluation Annual Report.

Opinion Dynamics Corporation 2013. Massachusetts Cross-Cutting Behavioral Program Impact Evaluation.

Reber, S. 2005. "Court-Ordered Desegregation: Success and Failures Integrating American Schools since Brown versus Board of Education," *The Journal of Human Resources* 40:559–590.

Rosenbaum, P.R., and D.B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects."

State and Local Energy Efficiency Action Network. 2012. *Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-based Energy Efficiency Programs.* Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman, Lawrence Berkeley National Laboratory.