

WHITE PAPER

Evaluating Opt-In Behavior Programs: Issues, Challenges, and Recommendations

California Public Utilities Commission – Energy Division

Report No.: CPU0088.01, Rev. Version 01

Date: July 31, 2014



Project name: White Paper DNV GL - Energy
 Report title: Evaluating Opt-In Behavior Programs: Issues, Challenges, and Recommendations [Office Post 2]
 Customer: California Public Utilities Commission – Energy Division [Office Post 3]
 Contact person: Ms. Valerie Richardson [Office Post 4]
 Date of issue: July 31, 2014 Tel: [+00 000 000 000]
 Project No.: [Enterprise No]
 Organization unit: Policy Advisory and Research, U.S.
 Report No.: Draft Version 01

Task and objective:

Prepared by:

Verified by:

Approved by:

Valerie Richardson, Principal Consultant
Consultant

[Name]
[title]

[Name]
[title]

Miriam Goldberg, Ph.D.,
Director and Country Manager, Sustainable
Energy Use

[Name]
[title]

[Name]
[title]

[Name]
[title]

-
- Unrestricted distribution (internal and external) Keywords:
 Unrestricted distribution within DNV GL [Keywords]
 Limited distribution within DNV GL after 3 years
 No distribution (confidential)
 Secret

Reference to part of this report which may lead to misinterpretation is not permissible.

Rev. No.	Date	Reason for Issue	Prepared by	Verified by	Approved by
0	[yyyy-mm-dd]	First issue	Valerie Richardson and Miriam Goldberg		



LEGAL NOTICE

This report was prepared under the auspices of the California Public Utilities Commission (CPUC). While sponsoring this work, the CPUC does not necessarily represent the views of the Commission or any of its employees except to the extent, if any, that it has formally been approved by the Commission at a public meeting. For information regarding any such action, communicate directly with the Commission at 505 Van Ness Avenue, San Francisco, California 94102. Neither the Commission nor the State of California, nor any officer, employee, or any of its contractors or subcontractors makes any warrant, express or implied, or assumes any legal liability whatsoever for the contents of this document.

Copyright © 2014, DNV GL.

This document, and the information contained herein, is the exclusive, confidential and proprietary property of DNV GL and is protected under the trade secret and copyright laws of the United States and other international laws, treaties and conventions. No part of this work may be disclosed to any third party or used, reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or by any information storage or retrieval system, without first receiving the express written permission of DNV GL. Except as otherwise noted, all trademarks appearing herein are proprietary to DNV GL.

Table of contents

1	EXECUTIVE SUMMARY	1
1.1	Goals of This Work	1
1.2	Key Conclusions and Recommendations	1
2	INTRODUCTION.....	3
2.1	What is a Behavior Program?	3
2.2	Approach	3
3	MEASURING CHANGE IN CONSUMPTION	5
3.1	Evaluation Based on Observed Changes in Energy Consumption	5
3.2	Consumption Data and Granularity	6
3.3	Measuring Change in Consumption	6
4	CONSUMPTION DATA ANALYSIS AND BEHAVIOR PROGRAMS	12
4.1	Self-Selection	13
4.2	Randomized Controlled Treatment (RCT)	19
4.3	Other Issues	20
5	LITERATURE REVIEW	26
5.1	Modeling Approaches	26
5.2	Highlights of Recent Work	30
6	CASE STUDY: PG&E PROGRESSIVE ENERGY AUDIT TOOL	35
7	APPROACHES NOT BASED ON CONSUMPTION DATA ANALYSIS FOR OPT-IN AND OPT-OUT....	36
7.1	Analysis of Explicit Action and Behavior Changes	36
7.2	Comparison Regions	37
8	CONCLUSIONS	38
8.1	Recognizing the Challenge	38
8.2	Recommendations	38
8.3	Improving Available Methods	39
9	CITATIONS	40

Table of Figures

Figure 3-1: Pre-Post Comparison without Confounding Effects	7
Figure 3-2: Pre-Post Comparison with Confounding Effects.....	7
Figure 3-3: Pre-Post Comparisons with Counteracting Confounding Effects.....	8
Figure 3-4: Pre-Post Comparisons with Confounding Trends	9
Figure 3-5: Matched Comparison, Post-Only.....	10
Figure 3-6: Matched Comparison, Difference of Difference Approach	10
Figure 4-1: Selection Effect Caused by a Time-Varying Characteristic.....	16
Table 4-1: Summary of Customer Groups with Combined Random Assignment and Opt-In Programs	22
Table 4-2: Average Change Components by Customer Group with Combined Random Assignment and Opt-In Programs	22
Table 4-3: Comparison Across Cells with Combined Random Assignment and Opt-In Programs	24



Table 4-4: Average Change Components by Customer Group with 2X2 Random Assignment 25
Table 4-5: Average Change Components by RCT Group Given Opt-In Program 25



1 EXECUTIVE SUMMARY

1.1 Goals of This Work

This paper began with the goal of providing initial evaluation methods for Pacific Gas & Electric's (PG&E) Progressive Energy Audit Tool (PEAT). More broadly, the paper assesses and recommends approaches for evaluation of the new generation of opt-in behavior programs. To that end, this work provides the following:

- An overview of the issues and challenges of evaluating opt-in behavior programs
- A summary and literature review of work done on these kinds of programs to date
- A preliminary look at the PEAT program as a concrete example of the application of these methods to this kind of program.
- A summary of this discussion in a set of recommendations relating to the evaluation of opt-in behavior programs.

For purposes of this paper, a Behavior program is one that attempts to influence customers to change their physical assets (energy-related investment behavior) and/or their operations (premise and dwelling use behavior) using information and encouragement methods, without directly providing financial assistance or tracking specific actions taken. These programs include audit-only programs, targeted information programs, and comparative information programs. They may include encouragement to participate in other programs that do include incentives and assistance for explicitly tracked measures, or to participate in upstream programs that don't track measures to customers.

1.2 Key Conclusions and Recommendations

1.2.1 Recognizing the Challenge

Recent behavioral 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 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 modeling, 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 behavioral 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.

1.2.2 Recommended Methods

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

- A combination of the variance- in-adoption model (VIA) method and matched comparison group should be used, depending on the specific characteristics of the program.
- VIA should be used provided that:
 - 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.
- Even with the above conditions met, inclusion of a matched comparison group with the VIA model should be tested as part of the analysis.
- For opt-in programs that start on a single date, a matched comparison group with weather normalization must be used without VIA.
- Matched comparison groups should be treated skeptically if there is a substantial portion of the participant group that has few good matches among the nonparticipants.
- To support the quantitative measurement of consumption effects, a qualitative analysis of program data should provide evidence of changes due to the program.
- Other programs' claims for "joint savings," if any, need to be subtracted from the consumption-based estimate of behavior program savings when assembling a total portfolio claim.

1.2.3 Improving Available 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 in this direction are the following:

- Improved matching algorithms based on key consumption parameters should be explored as part of whichever method is pursued.
- Existing RCT program data sets should be mined to better understand the extent of selection bias with particular analysis approaches.

1.2.4 Recommendations for PG&E's PEAT Program

The quantitative measurement of consumption change due to participation in a program ultimately relies on the correlation of the state change (e.g., installation of a widget) with a change in consumption levels. In the context of opt-in behavior programs, the primary piece of time-specific, state change information is the initial opt-in date. While web-based programs capture substantial amounts of site-level data, there are limited opportunities for these data to inform regression models as to when measurable consumption change might have occurred. In this respect, the PEAT Program is fundamentally similar to other opt-in programs. Aspects such as the specifics of the interface, the information offered, or the motivational structure may vary across programs, but the data available to support quantitative evaluations are quite limited. Thus, recommendations for PEAT would follow the recommendations above given the following starting point:

- Evaluation of PG&E's PEAT program should begin with participant analysis to assess whether a VIA or matched comparison group approach would be more appropriate.

2 INTRODUCTION

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

- An overview of the issues and challenges of evaluating opt-in behavior programs.
- A summary and literature review of work done on these kinds of programs to date.
- A preliminary look at Pacific Gas & Electric's (PG&E) Progressive Energy Audit Tool (PEAT) as a concrete example of this kind of program.
- A summary of this discussion in a set of recommendations relating to the evaluation of opt-in behavior 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 behavior 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. The PEAT is an example of the new generation of opt-in behavior programs that are being rolled out in California and other states.

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

2.1 What is a Behavior Program?

For purposes of this paper, a Behavior program is one that attempts to influence customers to change their physical assets (energy-related investment behavior) and/or their operations (premise and dwelling use behavior) using information and encouragement methods, without directly providing financial assistance or tracking specific actions taken. These programs include audit-only programs, targeted information programs, comparative information programs. They may include encouragement to participate in other programs that do include incentives and assistance for explicitly tracked measures, or to participate in upstream programs that don't track measures to customers.

2.2 Approach

The discussion of opt-in behavior programs tends to start with a discussion of randomized controlled trial (RCT) experimental design. RCT experimental design is a particular way of organizing a program that promotes the evaluation of that program. The theory and practice of RCT come from statistics and experimental sciences and has had applications in the area of energy programs. There are examples of RCT experimental designs in current behavior programs and this approach to these kinds of programs has played an important role in the increased stature of behavior programs in recent years.

Not all behavior programs, however, are designed as RCT experimental designs. There are good reasons for this. Opt-in programs are more flexible with respect to offerings, easier to target and as a result may generate more cost effective savings. Furthermore, they can be implemented with broad accessibility that is



consistent with rate-payer funded programs. There is a reasonable perception that the RCT design is limiting and perhaps unnecessary.

Rather than holding opt-in programs to the standard of an RCT program from the outset, we discuss the options for evaluating these kinds of program. We start with a discussion of measuring change in consumption for any program. We discuss the strengths and weakness of various methods for addressing the needs of an opt-in behavior program evaluation. This approach illustrates the range of challenges that face any evaluation measuring a change in consumption and the assumptions that may be required to get a result. This approach focuses on the assumptions that will be required to measure savings for opt-in behavior programs. In the process this discussion helps to illustrate why RCT experimental design is an elegant solution to the challenges.

The following sections focus on this challenge. Section 3 looks at the options available for measuring consumption change in general. Section 4 extends the discussion of measuring consumption changes to specifically focus on behavior programs. Section 5 introduces the common modeling approaches in the literature regarding opt-in behavior programs. In addition, we discuss some of the relevant publications on the subject. Section 6 provides a limited overview of the PEAT program and a discussion of the possible approaches to estimating savings for this program. Section 7 briefly discussed evaluation approaches not based on consumption data. Finally, Section 8 sums up the findings from the paper.

3 MEASURING CHANGE IN CONSUMPTION

3.1 Evaluation Based on Observed Changes in Energy Consumption

The measurement of the effect of behavior programs typically uses consumption data as the basis of the measurements. This is the general method explored in this White Paper. Alternative approaches are addressed briefly in Section 7.

Evaluation of behavior-based programs using analysis of consumption data is attractive for several reasons. One is that consumption data are readily available. Another is that pure behavioral changes (as opposed to equipment installation) are hard to verify, as noted. When changes are small, rare, and/or inconsistent, observed change across an affected population gives a form of ground truth, that's not dependent on isolating large numbers of small changes. This analysis also incorporates all program-induced effects, purely behavioral and physical, and their interactions. The key challenge is to isolate the program effects from other changes.

An additional motivation for focusing on consumption analysis is the recent RCT program designs. These designs have effectively used consumption data analysis with stronger grounding than has often been available with traditional programs. With this encouragement, practitioners are now re-examining use of consumption analysis with for programs with small average effects, while moving away from the constraints of RCT.

The use of consumption data to measure energy efficiency-related savings has a long history. The use of consumption data to measure the effect of behavior programs is usefully understood in the context of this history. This section focuses on the methods used to measure change in consumption using consumption data. Consumption data analysis, usually referred to as billing analysis, has been used in the energy evaluation field for more than 30 years.¹ During this time, the strengths, weaknesses, and unresolvable concerns of billing analysis have been well explored. While billing analysis is a complex area, there is a reasonable amount of consensus on the methods. Much of this consensus is discussed in the National Renewable Energy Laboratory's Universal Methods Project chapter on Residential Whole Building Retrofit with Billing Analysis (NREL 2013). While that chapter focuses on whole-building retrofit, there are, in fact, many similarities between measuring the effect of a behavior program and measuring the savings related to a whole building retrofit.

Billing analysis is traditionally used for whole-building retrofit evaluations because in measuring the overall change in consumption, the approach addresses the complex interaction effects of different measures. For instance, a billing analysis will measure the combined effect of a new efficient furnace and increased insulation. An aggregate consumption approach is appropriate because combined savings of these two measures are not distinct and additive. The alternative, an individually estimated, measure-based approach would face the challenge of appropriately de-rating the combined savings and would ignore any related behavioral effects.

¹ M. F. Fels, ed., *Energy and Buildings* (Special Issue Devoted to Measuring Energy Savings: The Scorekeeping Approach), 9, no.1&2 (February-May 1986).

M. F. Fels, K. Kissock, M.A. Marean, and C. Reynolds, *PRISM Advanced Version 1.0 Users' Guide* (Center for Energy and Environmental Studies, Princeton, New Jersey, January 1995).

M. F. Fels and K.M. Keating, "Measurement of Energy Savings from DSM Programs in U.S. Electric Utilities," *Annual Review of Energy and the Environment*, Annual Reviews, Inc., 18 (1993): 57-88.

Behavior program savings are also whole-building in nature and complexity. In addition to possible interactive effects, there is the sheer number and variety of potential savings actions that add up to behavior-related savings. Individual behavior-related savings are generally small, variable, and difficult to track. Only a consumption-based approach can capture the full, combined effect of these various activities.

3.2 Consumption Data and Granularity

Consumption data, as defined for this discussion, are measurements of the gas or electric consumption at a household or site over some period. Traditionally, monthly or bi-monthly data captured account/meter level consumption that provided the basis for billing customers for the energy used; thus, the term billing data. The data are generally maintained over multiple years, providing a historical record of site-level consumption. This source of accurate retrospective data has been well used by the evaluation industry.

The discussion here is framed mostly in terms of monthly consumption data. With the advent of advanced meter initiatives, data that are more granular are increasingly available. While these finer interval data offer some new opportunities, especially with regard to looking at peak load effects, it is not clear that these finer data intervals will alter the basic challenges of measuring change in consumption over time. After discussing the measurement consumption change in the context of monthly data, we will return to the full range of consumption data granularity and consider the opportunities therein.

3.3 Measuring Change in Consumption

A change in consumption caused by any program implies two time periods, one before the program interaction has occurred (pre) and one after (post). The transition time between the two periods might take many months for a building retrofit, or may occur on a single day with an entry into an online audit tool. Furthermore, the change in consumption may be immediate and consistent going forward or may develop gradually over time. Regardless, there is a time period prior to the start of the program activity that reflects pre-program normality and a period after program activity that includes the evidence of any change in consumption due to that program. The most basic measure of consumption change starts with a pre- to post-program comparison.

In many ways, the pre-post comparison is the basis for all approaches to measuring change in consumption. The apparent alternative to this same-site, pre-post comparison is a comparison with a non-program site during the post-period timeframe. At first, this appears to be a completely different approach but in practice, it is not. Comparison across sites only works if the two sites are otherwise similar and this similarity is invariably established using pre-program data. As a result, even the comparison group approach, which ostensibly foregoes the pre-post change in consumption, relies on it heavily.

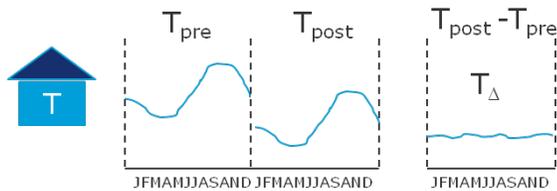
3.3.1 Pre-Post Differences

Billing analysis, at its most simple level, measures the difference (Δ) in consumption at a site between two periods. The goal is usually to use the measured consumption Δ as an estimate of the savings resulting from some program-related change at the site. For this to be the case at a single site, the household energy consumption would have to be identical between the two periods other than the program-related change.

Figure 3-1 shows an example of a site-level electric consumption load shape for a year before and after a program-related change. This example could represent, for example, the replacement of an old refrigerator with an energy efficient refrigerator. A simple, period-to-period delta provides the load shape of the

consumption Δ (T_{Δ}). Assuming the pre- and post-program consumption are effectively identical absent the reduction in consumption due to the refrigerator, the consumption Δ equals the savings load shape.

Figure 3-1: Pre-Post Comparison without Confounding Effects

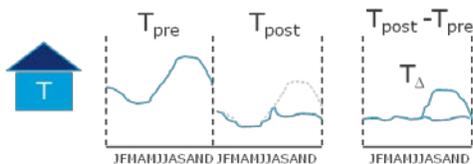


In actuality, many other things could cause a change in consumption between two time periods. A partial list would include:

- Different weather
 - Changes in square footage, occupancy, etc.
 - Changes in, or addition of, electricity-consuming equipment
 - Changes in behaviors related to energy consumption

All of these non-program variables have the potential to affect consumption levels in either the pre- or post-program period. If any of these period-to-period changes occur, they will confound the use of the measured consumption Δ as an estimate of consumption savings. A hot summer in the post-period could increase consumption in the post-period and make the consumption Δ smaller as a result. Alternatively, the air conditioner could break in the post period and the customer could decide to go without. This will decrease consumption in the post period and increase the consumption Δ . Figure 3-2 shows a post-program consumption as it might look with a broken AC. The resulting consumption Δ now reflects the actual savings and the confounding broken AC effect.

Figure 3-2: Pre-Post Comparison with Confounding Effects



These kinds of confounding effects will be present to some degree in all site-level, pre-post consumption Δ s. After all, while much may remain similar at a site from period to period, it is improbable the site-level consumption will stay identical apart from the program effect itself. These site-level confounding effects represent error in the direct use of the consumption Δ as an estimate of savings due to the program. The success of a billing analysis lies in its ability to control for as much of this non-program change as possible. Within a single site, or across a program population, many smaller, less easily identified, reported, or analyzed effects are happening all the time, in both directions. Some of these cancel each other out while some contribute to overall directional trends. The key is to find the program effect over and above these naturally occurring non-program changes.

Billing analysis has techniques that address some of these concerns. Weather normalization, for example, is designed to remove the differential effects of weather in the pre- and post-program periods. The models

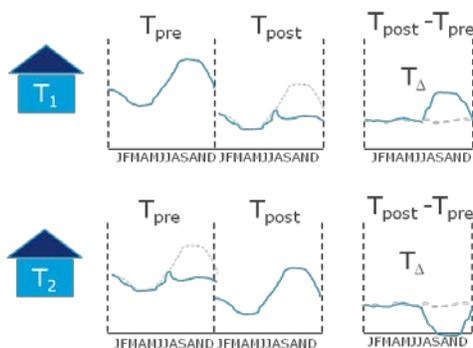
allow us to recalibrate pre- and post-program consumption as if they were produced by the same weather. This limits the potential for weather-related changes to confound the consumption Δ . Because of the wide availability of weather data and the strong correlation between heating and cooling consumption and degree days, this kind of weather normalization is feasible and is standard in most billing analysis approaches. These kinds of techniques are actually best applied at the site-level.

The other confounding factors are a different challenge to address than weather. They cannot be addressed directly at the site-level. The range of events that can affect consumption across time is effectively infinite. Data will never be available with which to control the full range of these other kinds of confounding effects. Instead, these factors have to be addressed at the aggregate level. This is why billing analysis is performed on large groups of sites, usually the whole population of a program. This allows us to consider specific site-level non-program change as part of a distribution of non-program change across the group.

In a simple form of analysis, we calculate an average weather-normalized consumption Δ across all sites in the group. This estimate is the average true savings plus the average of the site-level non-program changes. If the average non-program change across all of the sites equals zero, the estimate of average savings will be unbiased with respect to these non-program changes. An estimate is called “unbiased” if on average the estimate is expected to be equal to the true value of interest. That is, the average estimate is equal to the average true value, averaging over homes and over time. There are plenty of reasons, however, why the average non-program change might not equal zero. The non-zero, expected average non-program change is the bias of the estimate.

There are types of site-level consumption change that tend not to add to bias. These are temporary increases or decreases in consumption that will equal out if they show up equally in the pre- and post-program periods. For instance, if going without AC were only a single season occurrence, then a site choosing to go without AC in the pre-period would approximately balance the site going without AC in the post-period. The two sites represented in Figure 3-3 would approximately counteract each other’s effect. If the likelihood of going without AC is approximately the same through time and across the group of sites, then the average effect on consumption Δ should be near zero. If these kinds of non-program change occur on an approximately constant basis through time, they will not bias the end result. In the overall distribution, they represent noise (variation) but not bias.

Figure 3-3: Pre-Post Comparisons with Counteracting Confounding Effects

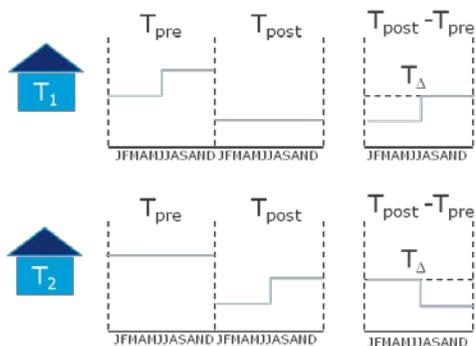


It is a different matter if there is a one-time, short-term shock related to a low-level natural disaster or some other system issue. These kinds of shocks tend to occur during a limited timeframe, affect a large

portion of the population, and affect consumption in a common direction. This kind of occurrence could lead to substantial bias of the savings estimate if not dealt with in the estimation method.

Finally, consistent increases or decreases in consumption will show up as a non-program difference in the pre-post consumption Δ whether they occur in the pre- or post-program period or both. Figure 3-4, shows two flat consumption load shapes that realize substantial savings due to a program. In addition to the program, each site also adds a motor that increases their otherwise flat load. One site adds the motor in the pre-period, the other in the post period. In both cases, the increase in load due to the motor will downwardly bias the estimate of savings from the program. The true reduction of consumption is flat and of the magnitude of the shift at the change point (at the dotted line level of T_{Δ}). The pre-post consumption Δ is lowered in both cases by the new motor consumption. That is, unlike the ongoing, short-term change discussed previously, a trend in consumption, if not addressed, will bias the estimate of savings regardless of when and how it occurs.

Figure 3-5: Pre-Post Comparisons with Confounding Trends



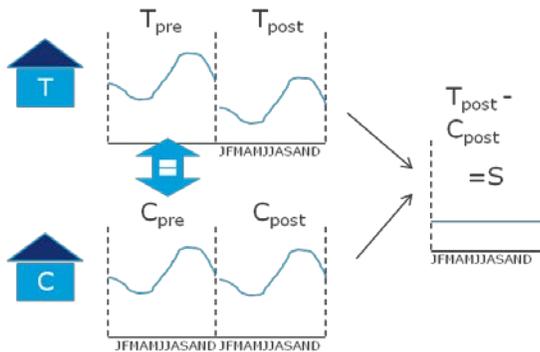
As with weather-normalization, there are billing analysis techniques that attempt to address these remaining confounding effects at the group level. Within the simple, pre-post consumption Δ scenario, a fixed-effect modeling approach can control for certain exogenous shocks and trends. The fixed-effects model estimates many program-participant pre-post comparisons, with a range of participation dates, in a single model. Instead of a single site-level, pre-post consumption Δ , the model measures an average consumption Δ across many sites that participate in the program at different times. Because only a subset of sites are moving from the pre- to post-program period at any given time, there are many sites to inform an estimate of non-program changes for each month. The approach can control for consumption effects that are similar across the whole population (e.g., a general trend or widespread single month aberration). These models control for effects on an average basis across the population. To the extent that systematic, non-program effects remain that are not accounted for in the model, these factors will tend to bias the estimated program effects.

3.3.2 Comparison Site Differences

An alternative to measuring change over time with a pre-post approach is to measure post-period change against a comparison site. This approach locates a site that is identical in every respect to the program site, but without the program interaction. In theory, it is possible to identify comparison sites using non-consumption related data. Then, instead of calculating a pre-post consumption Δ , we would calculate a treatment-comparison consumption Δ . To the extent that the comparison group was representative of that program/treatment site, the consumption Δ would equal true savings.

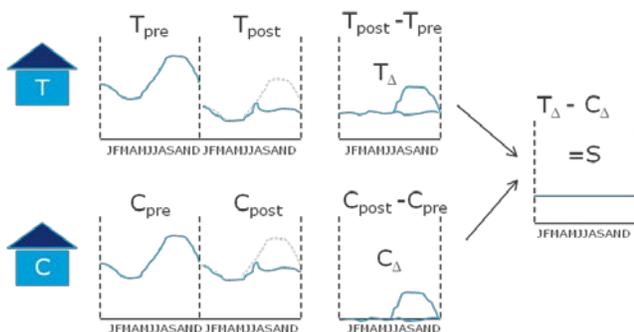
In practice, while there are increasing data options available for characterizing households beyond their consumption data, those additional data do a poor job of identifying households with similar energy consumption characteristics. As a result, the matched comparison approach generally uses pre-program consumption as the basis for choosing comparison sites. Figure 3-5 illustrates how this works. The approach uses pre-program consumption data for the program/treatment site (T_{pre}) to identify a comparison site with similar consumption in that pre-program time period (C_{pre}). Once matched, the post period, comparison site consumption (C_{post}) provides an estimate of the program household in the post period (T_{post}), but without the program change. If the sites were identical except for program interaction, the difference in the consumption of the two post-period households would equal savings.

Figure 3-6: Matched Comparison, Post-Only



Creating matched comparison groups based on pre-participation consumption has been a standard billing analysis approach for decades. In practice, instead of comparing only the post-period consumption data for the program and comparison households, evaluators use a difference of difference approach. This method effectively combines the pre-post and matched comparison approach as illustrated in Figure 3-6. The estimate of savings is the difference of two pre-post consumption Δ s. In effect, the program household pre-post consumption Δ is adjusted by the comparison household pre-post consumption Δ . Combining the two approaches reduces the effect of initial differences between the two groups, while controlling for non-program changes between pre and post.

Figure 3-7: Matched Comparison, Difference of Difference Approach



The difference of difference equation can be written $S = (T_{pre} - T_{post}) - (C_{pre} - C_{post})$. In this equation, the comparison group pre-post consumption Δ . In this formulation, the comparison can be seen as an estimate of the non-program related change that is occurring at the treatment site. If the comparison group Δ is a good estimate of these non-program changes, then its subtraction will produce an unbiased estimate of the treatment site savings. The result will be a good estimate if, on average, across the matched group of



comparison sites, the same trends and shorter term shocks are summarized in the pre-post error as are present for the treatment group, outside the program effect. This organization of the equation highlights one potential shortcoming of the comparison group approach. The approach assumes that given the pre-period similarity, the treatment and comparison groups will experience comparable non-program change between the two periods.

The equation can also be organized as $S = (C_{\text{post}} - T_{\text{post}}) - (C_{\text{pre}} - T_{\text{pre}})$. This organization of the equation emphasizes the post-period, comparison-treatment group consumption Δ . The pre-period, comparison-treatment group consumption Δ (the second parenthetical component) should be approximately zero if the matching is effective. This organization of the equation highlights the other potential shortcomings of the comparison group approach. This approach implies the pre-period difference between the groups is close to zero.

What does it mean that the pre-period delta is on average approximately zero?² The matching is done using minimum distance or propensity score algorithms using calendarized billing data. By construction, the average pre-program difference ought to be near zero. These monthly, site-level comparisons, however, mask a great deal of actual variation among sites. Monthly consumption values are complex combinations of many smaller loads. Matching on monthly consumption provides a comparison site of similar aggregate magnitude and shape across the pre-program year, but says little about the underlying loads and behaviors. Groups of sites, whose consumption look identical in aggregate, may, for instance, be the product of a range of combinations of size and energy intensity. If there is variable weather across the geographical area, then a further dimension may be simplified away with the matching algorithm. While the overall consumption might be similar, the underlying physical and human characteristics could be extremely different.

The ignored underlying variation in the pre-period becomes important when we recognize that the difference in difference structure assumes that similar non-program, pre-post changes affect the treatment and control groups across the periods. Returning to the first formulation of the equation, the comparison group's consumption change corrects for the treatment group's non-program change. Yet, to the extent that matching efforts are limited, the same must be said for resulting pre-post consumption Δ . If, for instance, a site with weather-correlated consumption is mistakenly matched with a similarly shaped but non-weather-correlated site, the effects of weather will clearly be different for the comparison site than the program site in their respective pre-post consumption Δ . The concern, of course, is that if the comparison group is not compensating for the bias, it is definitely adding the variability of the savings estimate and in extreme situation could even contribute to the overall bias of the estimate.

² Some evaluators actually drop the pre-period consumption Δ because it is, by construction, 0, at least for large enough samples of relatively homogeneous customers.

4 CONSUMPTION DATA ANALYSIS AND BEHAVIOR PROGRAMS

The above introduction provides a high-level summary of the issues that accompany the use of consumption data to measure the effect of a program. The discussion was deliberately general about the kind of program that was under consideration or even whether the sites were residential or commercial. The overarching message is that distinguishing true program-related savings within a pre-post consumption Δ 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 HVAC programs.

A primary reason the potential biases in a billing analysis result have been accepted for these kinds 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 underestimates. 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.

Behavior programs require a further level of scrutiny on the challenges of consumption data analysis. The most basic reason for this is the relative magnitude of expected behavior program savings. The potential bias becomes a much greater concern when the savings are less than 5%. As the potential but un-measurable 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 generate savings through the installation of measures. While human behavior may affect the exact level of savings generated by a measure, the behavior-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 behavior 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 trending and variable program effect from exogenous trends and weather variability.

In particular, opt-in behavior programs necessitate more scrutiny on the process of self-selection into a program and the implication for estimated effects. Self-selection is not a new issue or one that is confined to behavior programs, but like the more general potential bias issues associated with billing analysis, concerns related to selection, receded in the discussion surrounding the evaluation of measure-based programs as extensive early efforts failed to eliminate the problem. It was precisely the presence of RCT behavior programs that re-inserted the consideration of selection into the discussion of billing analysis. A primary reason for providing the high-level primer on the basic challenges of consumption data analysis in section 3.3 was to provide a context within which to understand, on a more intuitive basis, the issue of self-selection and the potential for resulting bias.

4.1 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” (p. 53, Imbens, et al, 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 understanding the mechanics and effects of self-selection more generally. It is possible to understand how self-selection causes trouble in simpler terms.

4.1.1 The Counterfactual

For the discussion of self-selection on simpler terms, we introduce the term “counterfactual.” The counterfactual for a site that has experienced a program intervention is that exact site in a parallel reality where everything is identical except for the program presence. The counterfactual cannot be observed, but the construct represents an opportunity to consider the range of possible realities that could exist in the absence of the program. It is a useful theoretical construct to contrast our attempts to find actual sites to serve as proxy counterfactuals.

Energy program evaluators use comparison groups to represent what happens in the absence of the program. When we construct a comparison group, we aspire to the counterfactual. The comparison group is a proxy counterfactual. A good proxy counterfactual will mirror the program site in all non-program-related changes. When we subtract the proxy counterfactual site’s consumption from the consumption of a site that has experienced program-related change, we isolate the effect of the program on consumption. The difference (or Δ) will include the possible subtle effects of the program without conflating non-program-related changes in consumption with the program-related changes.

In practice, the proxy counterfactual will always approximate the true counterfactual. Our ability to measure accurately behavior program savings depends on how well our comparison groups approximate the true counterfactual. The goal is to identify a group of proxy counterfactuals (or comparison group) that approximately mirrors the group of program participants. Regardless of the amount and/or quality of the data we use to characterize a comparison site, at the individual site level there will always be error relative to the true counterfactual. The goal is to develop the proxy group such that the resulting errors relative to the true counterfactual on average are approximately evenly distributed.

4.1.2 Self-Selection in the Group

Random assignment of a population to two groups produces two groups that are proxy counterfactuals of each other. If one group subsequently receives some treatment, then the untreated group provides a group proxy counterfactual for the treated group. The selection of the proxy occurs prior to any treatment or any decision to act on that treatment. Nothing about the subsequent treatment affects the assignment of individuals to either group.



Programs that follow an opt-in model changes the balance of group assignment. Participants receiving treatment choose to opt-in or “self-select” into the treatment group versus being randomly assigned. Once participants have opted into a program, the construction of a proxy using random assignment is no longer an option. The participant group is defined by the act of participating. Evaluators have tried to address this challenge by constructing a matched comparison group after the fact using available data to match the participants’ sites.

With good data, and a very large nonparticipant population, it may be possible to develop a comparison group that closely matches the participants prior to their participation on all of the observable characteristics. However, no matter how good this matching is there will remain one characteristic that is clearly unique to the participants—they chose to participate. We cannot observe the motivation to participate; and more importantly, we cannot observe if such motivation makes them more likely to make other choices such as changing consumption or installing energy efficiency measures.

Self-selection is the fundamental challenge to identifying a good comparison group after the fact. There are at least three challenges to identifying the proxy counterfactual sites for a matched comparison group. First, there is the question of what data are available to characterize the treatment site’s important characteristics so that a valid proxy can be identified. Second, assuming there are data available that allow characterizing the treatment site, does a site exists that matches the characterization well enough for the role of proxy counterfactual? Third, even if it is possible to find a match as proxy using energy consumption and general characteristics, we still have no knowledge on how much self-selection is playing a role on energy consumption after treatment to characterize properly what is counterfactual behavior without the program.

4.1.3 Identify the Proxy Counterfactual

The level of characterization that is required to identify a good proxy counterfactual is impossible to know. The goal is to capture all characteristics that could have an effect on the response to a program treatment. On the face of it, this seems like an impossible task. What makes this challenge potentially feasible is that we only have to succeed at this task on an average basis.

What level of characterization is required to identify a good proxy counterfactual? If, for example, building type were a sufficient basis, then any single-family dwelling could serve as a proxy for any other single-family dwelling. If the program participants were a random draw of 1,000 single-family dwellings, then a match group of single-family dwellings would be a reasonable set of proxy counterfactuals. If we pair the randomly drawn participants and nonparticipants based on no other matching or sorting criteria, the energy consumption differences would be substantial, both positive and negative. However, on average the differences between the two groups, apart from the program effect, would be approximately zero. That is, the nonparticipant households as a group would provide an unbiased representation of the randomly drawn participant households as a group absent the program.

Alternatively, consider a program that only attracts households with a swimming pool as participants. If the comparison group was still drawn from the general population of single-family dwellings, the outcome of the consumption comparison would be very different than in the prior example with the 1,000 randomly drawn participants. On average, houses with pools tend to be bigger and have higher energy consumption. This time, if we paired participants and nonparticipants in the post program period, the energy consumption



differences are still substantial and may still include both positive and negative differences.³ However, on average this time, the differences between the two groups, apart from the program effect, would *not* be approximately zero and the participant households would have higher consumption on average. The assumption that the two groups consumed the same amount of energy absent the program would be unfounded. The participants could lower their consumption substantially, but the comparison to the matched comparison group would not reveal it. The nonparticipant households as a group would provide a biased representation of the participant household consumption. Most importantly, this comparison group would be a poor choice for measuring program-induced savings among participants.

This is a simplistic example of self-selection causing selection bias. The households that opt into the program happen to be more likely to have pools. This characteristic may not be observable in the potential nonparticipants.⁴ More importantly, the basis for building the comparison group does not include this important characteristic. As a result, the comparison group is not populated with good proxy counterfactual households, and the estimate of difference between these two groups will be a combination of program participation and the effect of pool ownership.

4.1.4 Pre-Period Consumption as the Basis for the Proxy Counterfactual

As discussed in the previous sections on general consumption analysis, comparison groups are frequently selected based primarily on pre-program consumption. If two households are similar with respect to consumption during the present year, there is a good chance they will be similar in the next year. Put another way, consumption from one year is considered a good predictor for consumption the following year. If we randomly picked a group of single-family dwellings and then constructed a matched comparison group from the remaining population of single-family dwellings using pre-period consumption, we could expect the future consumption of the matched comparison group to provide an unbiased estimate of the future consumption of the random participant group.⁵ In fact, we would expect a comparison group constructed this way to provide a proxy counterfactual that is still unbiased while being less variable than the comparison group based on building type alone.

While pre-period consumption may be a more informative characteristic on which to base the proxy counterfactual than building type, it is still not a perfect estimator of future consumption. Individual households are constantly changing as the inhabitants change, and the physical characteristics of the house and end uses evolve. In aggregate, residential load trends up and down depending on the economy and other factors. The use of a comparison group based on pre-period consumption assumes that the two groups will change in similar ways over time. That is, the pre-period consumption for both groups does not have to be unbiased estimators of future consumption. They just need to be similarly biased estimators for the future consumption. The comparison group's pre-post relationship needs to be a good estimator of the participant group's pre-post relationship. If both participant and comparison groups drop consumption by 5% in the post-period because of an economic downturn, their pre-period consumption would also be a similarly bad estimator for the next period consumption. Despite this, the comparison group would still provide an unbiased estimate of participant consumption in the post period.

³ For now, we are just thinking about the post-program difference. The difference of difference approach addresses these concerns to some degree but this section is trying to illustrate the concept of selection.

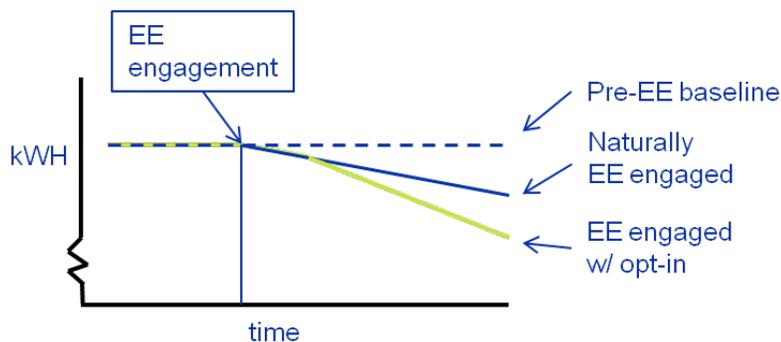
⁴ The characteristic may or may not be recognized in the participant group. Either way, if it is not accounted for in the nonparticipant group it may lead to biased estimates of difference.

⁵ This example assumes we are working with a very large dataset where we can match households in both groups by consumption level and not exhaust the available sample by consumption level (i.e., we have equal number of high consumption consumers, etc., for both groups.)

4.1.5 Self-Selection with Matched Comparison Groups

Self-selection can still be a concern when pre-period consumption is the primary basis for selecting the proxy counterfactual. In the prior example, the self-selection of households with swimming pools into the participant group undermined how well a comparison group constructed of other single-family dwellings estimated participant group consumption. Figure 4-1 illustrates a scenario where the selection process is correlated with the change in consumption. Because of the selection process, the pre-period consumption would not be a good estimator of post-period consumption regardless of the program. In this scenario, some non-program influence moves a customer toward engagement with energy efficiency. This non-program influence could be reading an effective article on climate change or an unusually high bill. If this customer does not have an opt-in behavior program available to her, then she will pursue her new interest with whatever success she can muster given the available resources. This person's consumption characteristics could change considerably without any input from the opt-in behavior program. The opt-in behavior program is an additional tool that this customer can enlist in her non-program motivated efforts to change her energy consumption. A good tool will facilitate her efforts, increasing the consumption reduction.

Figure 4-1: Selection Effect Caused by a Time-Varying Characteristic



The matched comparison household is only matched on pre-program consumption, not the newfound engagement in energy efficiency. Their pre-program consumption would be a poor estimator for the participant's post-period consumption in the absence of the program.

In terms of measuring the effectiveness of the program, the true counterfactual for an energy efficiency-engaged customer that has opted into the program would be an engaged customer without the assistance of the program's resources. By contrast, the proxy counterfactual comparison group constructed based on pre-period consumption characteristics would perpetuate the pre-period regime as the counterfactual. This would lead to crediting the full consumption reduction to the behavior program when in fact much of savings might have occurred in the absence of the program. In modeling terminology, opt-in-related savings would be confounded with natural engagement-related savings.

In Figure 4-1 the bias to the estimated savings would be upward. A scenario resulting in the opposite bias could also be constructed. Perhaps, for instance, homeowners who make energy-consuming additions to their homes might be disproportionately likely to opt into a program at that time as a kind of atonement for their increased energy consumption. If consumption increased, then it would be confounded with opt-in program related savings. The effect of this confoundedness would bias the savings estimate downward, correlating program participation with the increase in consumption.

4.1.6 Time-Invariant vs. Varying Characteristics

There is a distinction made between time-invariant and time-varying characteristics. Many aspects of consumption are, in fact, time-invariant to at least some degree. The physical characteristics of a house or site often do not change much over time. Human beings also have characteristics that may remain effectively un-changed over time. To the extent that consumption characteristics do not change over time (time-invariant), pre-period consumption is a particularly useful basis for constructing a comparison group. However, if consumption characteristics were completely time-invariant, then pre-period consumption would be a perfect estimator of post-period consumption and there would be no need for matched comparison groups. This would mean self-selection was a non-issue.

The prior example in Figure 4-1 provides one scenario where consumption is likely time varying. In fact, at the household or site level, it is hard not to consider some change in consumption as the norm. Families grow and shrink; the mix of end-uses changes; people get jobs and lose jobs. It is useful to think of it as a distribution of site-level pre-post consumption changes. Those changes may support an average trend that moves a couple percentage points. The site-level changes vary widely on either side of that average change. The key question is whether sites that opt into an opt-in behavior program look like a random draw from the distribution of pre-post consumption changes. This is impossible to test in the context of an opt-in behavioral program precisely because the consumption change would correlate with the program. At the same time, it does not make sense that voluntary participants would look like a random draw from the pool of all eligible customers.

4.1.7 Who is Available for the Comparison Group?

To this point, the discussion of self-selection has centered on characteristics and if we can appropriately identify households that could serve as proxy counterfactuals. The discussion has noted that certain characteristics may be unobservable and/or may change, and that those characteristics could be related to the targeted program outcomes, leading to possible selection bias in measured program effect. In this context, the challenge of identifying the perfect proxy counterfactual is just a limitation in the data. In theory, if we could track everything, including intentions, tenacity etc. then we could find a good match. This outcome leads us to the second aspect of self-selection. Once the selection process has occurred, are there any reasonable matches left in the population with which to populate the comparison group?

The classic free-ridership dilemma is a real-world example of this problem. To measure true net savings for a rebate program using consumption data alone, the evaluator would want to construct a proxy counterfactual comparison group that mirrored the participants buying decisions in the absence of the program. To do this, the comparison group should include the true proportion of households that would have purchased the energy-efficient version of the measure in the absence of the program. It is likely that a disproportionate number of this particular group of customers, who would purchase an efficient unit even without the program, would have taken advantage of the rebate windfall (we refer to these customers as free riders). This means that would-be, non-program efficient unit purchasers would be under-represented in the remaining group of non-program purchasers.

In this case, it's not just lack of information that causes the self-selection problem. In this case, the selection process is so effective, there may not be any decent proxy counterfactual remaining in the non-program population. It is worth noting that unobserved characteristics are still the root cause. In this example the key characteristic, willingness to pay full price for energy efficiency, is difficult to observe. In



this example, though, even if the characteristic could be observed, it is not clear how many would-be, non-program efficient purchasers would remain to be included in the comparison group.

4.1.8 Selection Bias Summary

This section describes what self-selection is, why the associated bias is real, and why some of things that claim to address the bias cannot be relied on.

- 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 out from the random selection leaves us with self-selection.
- Self-selection bias is a tendency of a particular savings estimate to be systematically over or understated due to self-selection. This bias exists when self-selection is present and analysis methods are used that are valid only if there is no self-selection, or if the methods used partially mitigate the effects of self-selection.
- Comparing participants' change in consumption with nonparticipants' is a common means of estimating program-attributable savings. This estimate has self-selection bias if:
 - the customers who join the program tend to be different from those who don't,
 - those differences are related to how their consumption would change even without the program, and
 - the comparison group isn't a good reflection of the kind of customers who would join the program.
- Typically, customers who join a program tend to be different from those who don't, and those differences are related to how their consumption would change even without the program.
 - For example; income, home characteristics, household composition, and energy attitudes will all affect the natural change in consumption in response to changes in the economy, prices, and weather. These same customer characteristics are also related to inclination to join a program.
 - Thus, using a simple comparison between participant and nonparticipant consumption change as an estimate of program savings will typically result in self-selection bias.
- That bias can be reduced to some extent by creating a matched comparison group, matching on observable characteristics.
- Also to some extent, matching on pre-program consumption captures the effects of interest from all those customer characteristics.
- There are three important limitations to the ability of any such matching to take care of self-selection:
 1. Whatever is used as a basis for matching, some factors or influences induced some customers to join and others not to, and the participant and comparison group are different in terms of those factors, usually unobservable. If those factors are also related to consumption changes, the comparison is subject to self-selection bias.
 2. Matching on pre-program consumption, or on the direct drivers of consumption, does not guarantee a good match on *change* in consumption absent the program.



As an illustration of both 1 and 2, customers may opt in to a program at the time they want to make improvements to their homes. Thus, the customers are similar up to the point they opt in, and that is the point at which their consumption would diverge even without the program. No test of similarity up to the point they join the program can reveal this.

3. If the program is sufficiently attractive to certain customer segments, there may be very few from those segments available as matches among the nonparticipants. Moreover, those who remain are potentially very unlike the participants in some unobserved ways that might be related to how their consumption changes.

4.2 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.

All of the consumption dynamics that we have been discussing continue to occur in both the treatment and the control groups. The activities that could lead to selection bias in a non-RCT context also continue. There will still be a complicated chain of causality between program and non-program activities and their ultimate effect on consumption. However, by construction, the two randomly assigned populations will experience, on average, all of these effects in the same way except for the program-related effects.

This point is well illustrated by both the standard, opt-out home energy report (HER) programs and the quite similar random encouragement design approaches. In both cases, the selection process occurs within the treatment group alone. In HER programs, people choose to read and perhaps act on the reports. Those households that do opt to take action have theoretical counterfactual households in the control group. Many households, perhaps the majority, remain completely unchanged by their interactions with the program. Those households also have theoretical counterfactual households in the control group.

RCT experimental design does come with its associate challenges. It requires that a set of sites are removed from the program process. This means that programs must be targeted with respect to publicity and information, but also with respect to access to the program benefits. This explains why RCT has been used primarily for behavioral programs. For these programs, the “benefit” (e.g., consumption history, information, and neighbor comparisons) that is denied to control group members is minimal, and the promotion is targeted to the site-level.

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.

4.3 Other Issues

4.3.1 Potential Double Counting

A standard piece of all opt-out behavior program evaluations is addressing the potential for overlap with other utility programs. Various referred to as channeling, uplift, and joint savings, etc., this analysis identifies increased uptake of other utility programs as result of participating in the behavior program. It is one of the goals of these behavior programs to promote this kind of additional participation, and in most instances, there is evidence that this uplift does occur. However, it is necessary to quantify the uplift so that the effect of incremental program participation is counted only once in a utility's total portfolio claim. Arguably, both the behavior program and the other program contributed to the savings that occurred, but both cannot account these savings.

Downstream rebate programs are carefully tracked so analyzing site-level participation in these programs is tractable. Upstream programs are similar in theory but are more challenging practically speaking. This discussion will focus on rebate program savings first and then discuss the upstream savings challenge afterwards.

The problem with jointly motivated savings is one of attribution. Any savings related to an increase in downstream rebate programs will be captured in the behavior program savings estimates. That is, from the perspective of the behavior program savings regressions, there is no distinction if the savings being measured are due to pure behavioral response or the installation of a rebated furnace that would not have been installed without the behavior program. On the other hand, these rebate savings are inevitably claimed through the rebate programs themselves. Regardless of which program will claim the savings, it is essential to quantify to amount of uplift/overlap savings, and assure that only one program (rebate or behavior) includes those savings in their claim. The quantification process is part of the behavior-program impact evaluation.

In an RCT experimental design, measuring the differential uptake of rebate program savings is straightforward. Rebate program savings are aggregated for the treatment and control groups and compared on an average site-level basis. Any increase in the treatment-group rebate savings relative to the control-group rebate savings represents an unbiased estimate of joint savings. As with the overall behavior program savings estimates, the RCT experimental design is fundamental to making these estimates of joint savings unbiased.

In the context of an opt-in program, this issue of joint savings remains but is complicated by the lack of a valid RCT control group. Two approaches have been explored for this problem. The most obvious simply uses the constructed comparison group as if it were a true control group. Just as the comparison group is constructed to be as close to the ideal counterfactual of the treatment group with respect to overall consumption change, it should be similarly useful as the counterfactual with respect to other program participation. Any concerns regarding potential bias of the overall savings estimates, model (or comparison



group construction) bias or selection bias, pertain to the joint savings as well. The only comfort lies in the fact that joint savings tend to be a relative small fraction the estimated behavior program savings.

It is worth noting that it is possible to bracket the degree of potential double counting. The rebate program savings generated by treatment group households alone is the maximum potential joint savings. This approach assumes that all rebate program savings were motivated by the behavior program. This approach also implies that the baseline participation in rebate programs is non-existent which is likely incorrect. However, this approach allows us to compare estimated behavior programs savings, or maximum joint savings, to the overall savings estimated from the consumption data.

4.3.2 Mix Opt-In/Opt-Out Programs

Opt-in programs are frequently combined with RCT experimental design in a mixed program design. The purpose of this mixed approach appears to be to leverage the advantages of the RCT design while still pursuing the opt-in strategy. In reality, this approach does little to facilitate the evaluation of the non-RCT portion of the program and considerably complicates the overall understanding of the program effects.

A common mixed design sends a form of home energy report to a randomly assigned treatment group while all households receive general publicity related to an available web portal. These reports may be simple and amount to targeted marketing for the online, opt-in option that is the primary program offering.

Alternatively, the home energy report may be the primary focus of the program with an opt-in web portal made available to a wider audience. In both cases, both components may generate savings and the evaluation should account all of these savings. For ease of discussion, we will refer to the randomly assigned treatment (home energy report) as a booster program, recognizing that one of the purposes of the program is to boost opt-in participation. The important characteristic of the booster program is that it is randomly assigned rather than chosen or opt-in.

Table 4-1. summarizes the make-up of the four groups created by this kind of combined program. As the table illustrates, the kinds of households in each column differ depending on whether they were exposed to the booster program.

- The natural opt-outs are in the booster program control group, so are not exposed to additional boosting. The natural opt-outs do not include households that opted in despite not receiving additional boosting (natural opt-ins). Natural opt-outs do include households that would have opted-in had they received the booster program (boosted opt-ins). Firm opt-outs are exposed to the booster program. As a result, they will not include households that are boosted to opt-in. Firm opt-outs will include only households that continue to opt out even when exposed to the booster program.
- The opt-in households in the control group will only include those willing to opt in without the additional encouragement of the booster program. The treatment group opt-in households will include all natural opt-ins along with those boosted opt-ins that needed the booster program to motivate opting into the voluntary program.

Table 4-2: Summary of Customer Groups with Combined Random Assignment and Opt-In Programs

Customer Population	Voluntary Program (Opt-In)		
	Group	Opt-Out	Opt-In
Booster Program (Random Assignment)	Control	Natural Opt-Outs	Natural Opt-Ins
	Treatment	Firm Opt-Outs	Natural Opt-ins + Boosted Opt-Ins

This variable composition of the households in each cell is combined with expected program effects on consumption that are also different across the cells. Table 4-3 summarizes the average pre-post period consumption change for each of the customer groups. The components of average consumption changes are:

- Natural change is what would have occurred with no program.
- Voluntary program (VP) change is the change in consumption due to the voluntary program alone, for those who opt in with no additional encouragement.
- Booster change is the change in consumption due to the Booster program alone (e.g. Change due to reports without any complication of an opt-in option)
- VP*Booster change is the interactive effect of the VP program and the Booster Program combined. The interactive effect could represent additional savings over the two effects individually as a result of synergy or could represent a reduction of savings if there is a degree of redundancy between the two separate effects.

Table 4-4: Average Change Components by Customer Group with Combined Random Assignment and Opt-In Programs

Customer Population	Voluntary Program (Opt-In)		
	Group	Opt Out	Opt in
Booster Program (Random Assignment)	Control	HHs: Natural Opt-Outs Pre-Post Δ : Nat. Δ	HHs: Natural Opt-ins Pre-Post Δ : Nat. Δ + VP Δ
	Treatment	HHs: Firm Opt Outs Pre-Post Δ : Nat. Δ + Booster Δ	HHs: Natural Opt-ins + Boosted-Ins ⁶ Pre-Post Δ : Nat. Δ + Booster Δ + VP Δ + VP*Booster Δ

⁶ Natural opt-ins and boosted-ins represent different self-selected segments of the population and, as a result, can be expected to have different pre-post Δ s. That goes for both natural Δ s, different as a result of different household characteristics, as well as VP or booster Δ s, as those different households interact with the different programs.



Because the composition of households that make-up each cell is not comparable, the average change components in each cell are not comparable to the components with the same name in other cells. For instance, the pre- to post-period natural change (nat. Δ) within each cell could be of different magnitudes for each cell. Similarly, we would not expect households that were unwilling to opt into a voluntary program even with exposure to a booster program to respond to the booster program in the same way as opt-in households (whether natural or boosted in).

These differences are crucial, because it is tempting to consider the difference between opt-out households with and without the booster program as a measure of the booster program effect. First, we cannot be confident that the natural change of the natural opt-in households appropriately accounts for natural change for the different mix among firm opt-out households. Moreover, the savings from the booster program for this group would only represent savings for firm opt-outs—households that opted out of the voluntary program despite the booster program. The firm opt-out booster program savings would likely differ from the booster program savings for either natural opt-ins or boosted opt-ins.

Table 4-3 summarizes the results of this kind of comparison across cells. In each case, a comparison that appears to have the potential to isolate an important component, instead gives a more qualified and less useful result. All of the differences are undermined by consumption effects that reflect the unique composition of the households self-selected into the groups. The simplest example of this is the fact that natural opt-out non-program change is not useful as a proxy for non-program change from any other cell. More importantly, the disentangling of the combined effects of the voluntary and booster programs is effectively impossible.

Table 4-6: Comparison Across Cells with Combined Random Assignment and Opt-In Programs

Customer Population	Voluntary Program (Opt-In)		(Opt-In) - (Opt-Out)	
	Group	Opt-Out		Opt-In
Booster Program (Random Assignment)	Control	HHs: Natural Opt-Outs Pre-Post Δ: Nat. Δ	HHs: Natural Opt-ins Pre-Post Δ: Nat. Δ + VP Δ	Natural opt-in VP Δ + (Natural Opt-in Nat. Δ - Natural Opt-out Nat. Δ)
	Treatment	HHs: Firm Opt-Outs Pre-Post Δ: Nat. Δ + Booster Δ	HHs: Natural Opt-ins + Boosted Opt-Ins Pre-Post Δ: Nat. Δ + Booster Δ + VP Δ + VP*Booster Δ	[Avg(Natural, Boosted opt-in Booster Δ) - Firm Opt-out Booster Δ] + [Avg(Natural, Boosted opt-in nat. Δ) - Firm Opt-out nat. Δ] + Avg(Natural, Boosted opt-in VP Δ) + Avg(Natural, Boosted opt-in VP*Booster Δ)
T-C		Firm opt-out Booster Δ + (Firm Opt-out Nat. Δ - Natural Opt-out Nat. Δ)	[Avg(Natural, Boosted opt-in VP Δ) - Natural Opt-in VP Δ] + [Avg(Natural, Boosted opt-in nat. Δ) - Natural Opt-in nat. Δ] + Avg(Natural, Boosted opt-in Booster Δ) + Avg(Natural, Boosted opt-in VP*Booster Δ)	

For comparison sake, if the 2x2 experimental design was actually randomly assigned for both treatments, then comparing the differences across the four cells would be relatively more simple. The components identified in each cell (natural Δ, etc.) would be comparable. Table 4-4 summarizes all the differences if all of the components are assumed comparable across cells.

Table 4-7: Average Change Components by Customer Group with 2X2 Random Assignment

Customer Population	Treatment #2 (Random Assignment)		T-C	
	Group	Control		Treatment
Treatment #1 (Random Assignment)	Control	Nat. Δ	Nat. Δ + T#2 Δ	T#2 Δ
	Treatment	Nat. Δ + T#1 Δ	Nat. Δ + T#1 Δ + T#2 Δ + T#1 * T#2 Δ	T#2 Δ + T#1 * T#2 Δ
T-C		T#1 Δ	+ T#1 Δ + T#1 * T#2 Δ	T#1 * T#2 Δ

The practical solution when the voluntary program is a voluntary, opt-in program is to compartmentalize the evaluations. The RCT (booster) portion of the evaluation provides an unbiased estimate of the savings given the presence of the opt-in program. Table 4-5 illustrates the combined groups and the different change components. Most importantly, the households in the two cells are not defined by their decision to opt in or not. They are randomly assigned, and as a result, comparable with respect to known and unknown characteristics apart from the treatment. The natural change and non-booster program voluntary program change will be removed from the treatment group change. This is a valid, unbiased estimate of these remaining components of the design. This result, however, does not account for voluntary program change (without boosting). Moreover, the forgone average voluntary change is relevant for the full RCT population, not just the treatment group.

Table 4-9: Average Change Components by RCT Group Given Opt-In Program

Customer Population	Voluntary Program Ignored	
	Group	Opt Out +Opt In
Booster Program (Random Assignment)	Control	HHs: Randomly assigned eligible Pre-Post Δ : Nat. Δ + VP Δ
	Treatment	HHs: Randomly assigned eligible Pre-Post Δ : Nat. Δ + Booster Δ + VP Δ + VP*Booster Δ
T-C		Booster Δ + VP*Booster Δ

The second step to this approach performs an appropriate opt-in program approach on the control group alone to estimate Booster control group VP Δ . As discussed, these estimates of savings have potential bias issues, but they do provide a rough estimate of the opt-in-related savings in the control group. For complete savings for the combined program, this estimate of opt-in savings must be included with the RCT savings estimate for both the missed treatment group savings and for the control group savings.

5 LITERATURE REVIEW

The overview of traditional billing analysis approaches to measuring consumption change provides a context within which to understand the range of models that have been used to estimate behavior programs. In this section, we will focus on models used to estimate savings for opt-in behavioral programs. Almost all evaluators use a standard fixed-effects model to estimate savings for RCT HER Programs. The presence of the RCT design makes it possible to use only the post-only comparison of treatment and control groups, but the standard approach uses the difference of differences fixed-effect regression.

5.1 Modeling Approaches

There are three primary approaches to modeling opt-in behavior programs that have been used to estimate opt-in behavior program savings. In this section, we discuss the model structures. In the following section we discuss the articles and reports where these models were used or discussed.

5.1.1 Variance in Adoption

One of the more common model structures used for estimating the effects of an opt-in behavior program, is called the variance in adoption (VIA). The model was used in the primary academic study that focused on these kinds of programs (Harding). It was also highlighted in the SEEACTION handbook on the evaluation of Behavior programs (SEEACTION), used for some program evaluations (Massachusetts) and discussed in conference papers (Provencher).

The VIA is structurally identical to the standard pre-post, pooled billing analysis approach that is used for a whole building retrofit. This pooled approach (generally discussed in Section 3.3) measures pre-post differences without a comparison group. There are only two differences with a standard pooled billing analysis model. First, instead of measuring an average post-period difference, the model measures consumption change as a function of the number of months since opting in. This is explained as accommodating the changing level of savings in the post-period of a behavioral program. Second, the model does not include weather variables. That is, there is no effort to weather normalize the consumption.

The model form is a simple fixed-effects model.⁷

$$E_{jm} = \mu_j + \phi_m + \gamma_p P_{jp} + \varepsilon_{jm}$$

Where

E_{jm}	=	Average daily energy consumption for site j and time interval m
μ_j	=	Unique intercept for each site j
ϕ_m	=	0/1 Indicator for each time interval m , time series component that track systematic change over time
P_{jp}	=	0/1 Indicator variable for each month p months after household j opted into the program.
γ_p	=	Change in consumption in month p after opting into the program

⁷ An alternative model that is commonly seen in the literature but is outside of the scope of this paper, is the use of log-transformed consumption as the dependent variable. In general terms, these models should give similar results.

ε_{jm} = Regression residual.

The model controls for a site-specific consumption effect as well as a month-year effect that will capture monthly variability in consumption across all sites given the other parameters. As with the standard whole building retrofit model, participants enter into the program throughout the evaluation period. The monthly post-period indicator variables identify consumption change at sites the same number of months removed from participation.

The advantage of the monthly (since opt in) savings estimates is that it makes it easier to get an estimate that is specific to a particular period. In contrast, measure-based savings estimates generally assume constant savings once installed. For this reason, models for measure-based programs general specify a single average annual savings parameter from the post period rather than monthly savings.

An addition is made to the equation in some evaluations that is used as a test of the validity of this model. A second set of parameters are included that mirror the savings parameters, P_{jp} , but characterize change for each pre-opt-in month. For example, a parameter would estimate any average differences across all sites in the third month prior to each site's opt-in date. Structured in a mirror fashion to the treatment effects parameters, these parameters are expected to not identify statistically significant change in the pre-period where no program effect is present. A finding counter to this could be used to justify a different approach.

The lack of weather variables is a more striking aspect of this model. This means, for example, that the first month savings estimate is an average of first month savings regardless of time of year a site opted into the program. Seasonal variability in the true underlying savings will negatively affect the precision of these estimates and mask meaningful differences in savings characteristics across the year.⁸ More problematically, the lack of weather variables implies that the monthly fixed effects will capture all weather effects in addition to trends and shocks without confounding them with the program savings that are expected to be variable and may trend over time. Finally, regardless if the savings estimate correctly captures the underlying savings, the weather-correlated aspect of the savings will not be captured in a way that can be put in "typical weather" terms.

More generally, this model recapitulates the assumptions discussed in Section 3.3.1 above about pre-post consumption models. This approach effectively assumes that confounding effects will on average not bias the result given the fixed effects in the model. The model is also unable to address the fundamental concerns regarding self-selection, which were discussed in section 4.1.

Weather could be incorporated into the VIA model.

$$E_{jm} = \mu_j + \phi_m + \beta_H H_{jm} + \beta_C C_{jm} + \gamma_P P_{jp} + \gamma_H H_{jm} P_{jp} + \gamma_C C_{jm} P_{jp} + \varepsilon_{jm}$$

Where

H_{jm}, C_{jm} = Average daily heating and cooling degree days for billing period m for household j.

β_H, β_C = Estimated heating and cooling trend in the pre-program period.

γ_H, γ_C = Estimated change in heating and cooling trend due to program participation.

⁸ As stated by Harding, "The underlying assumption is that, conditional on time-invariant household characteristics and aggregate month-specific shocks, all households that are k months away from enrolling in the offset program are identical (in expectation)." From M. Harding and A. Hsiaw, "Goal Setting and Energy Conservation," *Journal of Economic Behavior and Organization*, under revision (2013): 10.

This model captures a pre-period, weather-correlated consumption baseline and also measures the weather-correlated effect for each post-month estimate of savings. The original indicator variable savings parameters are still included along with the weather-correlated savings parameters. This model will do a better job of controlling for the potential weather-related bias. It will also give distinct estimates of program-related heating and cooling savings allowing for an estimate of savings under typical weather conditions.

5.1.2 Match Comparison Groups

The other common approach to evaluating opt-in behavior programs involves matched comparison groups. Match comparison groups are generally modeled in a difference of difference regression specification.

$$E_{jm} = \mu_j + \phi_m + \gamma P_m + \alpha T_j + \delta P_m T_j + \varepsilon_{jm}$$

Where

P_m	=	0/1 Indicator variable indicating the post-period.
T_j	=	0/1 Indicator variable indicating whether household j is in the treatment group.
γ	=	Estimated change in consumption in post period across both treatment and comparison groups.
α	=	Estimated difference in consumption between treatment and comparison groups in the pre-period.
δ	=	Estimated change in consumption in post period for the treatment group alone.

This specification recreates the difference of differences structure but in a regression framework. This structure as written, estimates the baseline consumption for the pre-period comparison group. It controls for treatment-group consumption differences across all periods and post-period consumption change across all sites. The remaining consumption change for the treatment group in the post-period is analogous to a difference of difference result.

In practice, the treatment group parameter is absorbed into site-level fixed effects. If the program participation starts at a single point, both the treatment group parameter and the post-period parameters will be absorbed into the site-level and month-year fixed effects, respectively. As the start point expands to multiple time-periods, the treatment and post-period effects become separately identifiable, in a statistical sense. These parameters will provide average annual estimates of the difference between treatment and control groups (controlling for those average differences), and the difference in consumption across all sites in the post period (trend), if the start of program participation is approximately evenly spread across a year.

5.1.2.1 Matching Algorithms

Various matching algorithms are proposed for developing comparison groups. The two primary approaches, minimum distance algorithms and propensity model matching both compare sites on a combination of site-level characteristics. Pre-period monthly consumption data are the primary inputs with additional characteristics incorporated either through stratification or directly in the propensity model. Some initial comparisons of the approach have been reported (IEPEC matching paper) but the conclusions in that paper are limited to the populations on which those evaluations were performed. At this point, it is too early to determine a clear winner between those two approaches.

In general, it appears that matched comparison groups are going to experience a renaissance. Whereas traditional billing-analysis comparison groups matched on annual consumption alone, current algorithms already incorporate substantially more information. The coming years are likely to see a great deal of experimentation on the best approaches for this task. There are obvious opportunities to pursue. Site-level regression models summarize site-level consumption data in models specifications with meaningful underlying structures. The model results would appear to provide an even better basis for site-to-site comparison than raw, inevitably calendarized consumption data.

5.1.3 Post-Only Model

A third approach uses a model that includes both treatment and comparison and pre- and post-program data, but structures the relationships to model only the post-period. The equation is written:

$$E_{jm} = \mu_j + \delta T_{jm} + \alpha E_{j(m-12)} + \varepsilon_{jm}$$

Where the only new component is

$$E_{j(m-12)} = \text{Average daily energy consumption for site } j \text{ and time interval } m-12, \text{ the same month but a year prior.}$$

For this model, prior year consumption for the same month is the primary explanatory variable. The monthly treatment parameter captures the average treatment-group difference for each month. The structure of this model is different than the difference of difference model and the determination of the exact implications are beyond the scope of this review. A comparison of results from this model compared with a more standard difference of difference would be useful.

5.1.3.1 Regression-Corrected Model

The Massachusetts Cross-cutting Behavioral Program Evaluation Integrated Report from 2013 (hereafter, MA evaluation), uses a similar model to this post-only model for what they refer to as a regression correction and that comes from an article by Abadie and Imbens (2011). The approach, as reported in the MA evaluation, compares treatment group average monthly consumption with a regression corrected version of the comparison group average monthly consumption. The correction is based on this simplified version of the above model estimated on the comparison group alone.

$$E_{jm} = \mu_j + \alpha E_{j(m-12)} + \varepsilon_{jm}$$

Using these estimated regression parameters that characterize the comparison group, the corrected comparison group average monthly consumption is calculated by applying the average monthly energy consumption of the treatment group. The argument is that this approach combines the strengths of the comparison group while using the regression correction to put the treatment and control estimates on the same basis with respect to distributions of the supporting data. This approach is relatively new in the literature but deserves to be further explored.

5.1.3.2 Combined Estimation of Savings and Joint Savings

It's worth noting that the actual version of these equations that were used for the MA report included an additional component to capture the effect of other rebate program effects. The equation was:

$$E_{jm} = \mu_j + \delta T_{jm} + \alpha E_{j(m-12)} + \phi_k O_k + \varepsilon_{jm}$$

Where

$$O_k = \text{A 0/1 indicator variable, 1 after the installation of energy efficiency program measure } k, \text{ 0 prior to that.}$$

φ_k = The average consumption effect of the installed measure k .

This approach includes the variables to capture post-period consumption change for the treatment group while adding variables that correlate to the timing of any rebated measure installation. This specification is a specific example of an approach, that has also recently appeared in RCT program evaluations, that claims to control for potential double counting directly in the savings equations, thereby producing an estimate of behavior program savings that is net of any joint savings. The approach appears to have some limitations with respect to both theoretical intent and an additional kind of selection bias.

This is a relatively new approach and there has been limited time to vet the details. There are two concerns that bear consideration. First, from a theoretical perspective, this approach claims to measure actual savings associated with rebate program participation. Strictly speaking, the issue of joint savings, however, is one of claimed savings rather than actual savings. The appropriate measure of joint savings (as measured by the RCT program approach) is a measure of uplift in claimed savings. From a double counting perspective, it should be the measure of savings on which the IOUs are paid or credited (net or gross). To the extent that actual savings as controlled for in the savings regression are less than the claimed uplift savings that are measured in the RCT context, then some degree of double counting remains. Alternatively, if actual savings are greater than claimed, the regression over-adjusts for double counting. This issue may be a relatively minor detail but it does point to issues related to the interaction of behavioral programs with other programs that will need to be addressed.

The second concern regarding the inclusion of rebate savings in the overall model specification is whether it actually succeeds in controlling for rebate-program-related savings as intended. There are two issues:

- First, this approach controls for average program participation across treatment and control. This is not the goal of the process. The goal is to control for the marginal difference between the treatment and control groups. At a minimum, the rebate program variables should be interacted with the treatment variable to estimate the marginal difference.
- Second, the rebate program variables function as independent variables in the regression but are dependent of the program outcome. In fact, what we are attempting to control is precisely the relationship between the decision to participate in a rebate program and the participation in the behavior program. The regression-based estimate of program effects is going to include savings that are specifically related only to the behavior program. For example, if a behavior program consistently reduces the tendency to “take back” (using more energy and therefore taking back savings as greater comfort) when an efficient furnace is installed by program participants, those furnaces will generate more savings than in nonparticipant homes where take-back occurs. These particular additional savings are due solely to the behavior program. This specification does not appear designed to capture that distinction. This is another instance of selection bias, and, with this kind of specification, it is even a problem with an RCT experimental design.

5.2 Highlights of Recent Work

The literature that informs the impact evaluation of behavior programs come from varied sources. The academic literature provides substantial guidance on the use of statistics or econometrics for the evaluation of program effects.

5.2.1 Recent Developments in the Econometrics of Program Evaluation

These include, in particular, a recent survey called Recent Developments in the Econometrics of Program Evaluation by Imbens and Wooldridge that was published in 2009. While this work does not focus on energy-related programs, the problems they address are exactly the problems facing the energy program evaluators with respect to opt-in behavior programs. The primary value of this survey of recent work is its clear statement about the limited options for addressing confoundedness, the term they use for self-selection. What is striking about the article is the extent to which just developing a matched comparison group on observable and time-invariant characteristics is challenging.

5.2.2 Goal-Setting and Energy Conservation

In addition, one academic paper analyzes an early opt-in behavior program. This paper has provided some of the guidance for subsequent evaluation work. Harding and Hsiaw looked at an early opt-in program that purported to motivate savings by having customers set goals. Much of the paper is spent developing the theoretical structure within which goal setting would be expected to be an effective motivation toward something like energy conservation. This aspect of the paper is somewhat puzzling to a reader familiar with what amounts to “goal setting” in this kind of opt-in behavior program. Programs that use this approach (including the Progressive Energy Audit Tool) give those who opt in the opportunity to respond to the provided tips. One of the responses is “Will do.” This response appears to be the basis for Harding and Hsiaw’s hypotheses about goal setting.

Fortunately, the actual quantitative measurement of program-related change is somewhat disconnected from the theory of goal setting. Harding used the model that is presented above as the VIA model. It is on this recommendation that this model has been attempted in most subsequent evaluation. In addition, Harding provides results for an alternative model that involves a propensity weighting designed to control potential selection bias. The documentation is insufficient to determine exactly what additions were made to the approach. No subsequent evaluation has attempted this approach.

In the program evaluation area there are a handful of public evaluations that make a real effort to bring the latest ideas in their analysis. In addition, these evaluations have been summarized and presented in various forms at conferences such as the International Energy Program Evaluation Conference.

5.2.3 Massachusetts Three-Year, Cross-Cutting Behavioral Program Evaluation

The MA evaluation includes two programs that can be considered opt-in programs. The WMECO C3Energy opt-in behavior program has been evaluated annually for three years likely making it the most evaluated opt-in program in the country. ODC/Navigant has been evaluating this program for those three years and they apply most of the methods already discussed here. A number of specific issues should be highlighted.

5.2.3.1 Mixed Program Design

The overall program design for the WMECO program is a mix of an opt-out informational mailer and an opt-in web portal.⁹ As discussed in section 4.3.2, this combination of experimental designs creates a complicated overarching analysis of savings. Navigant generally follows the recommended approach discussed in section 4.3.2. They measure the RCT experimental design savings and then adjust it with savings estimates derived from the control group opt-in households. The discussion in the report is not

⁹ The program also had additional waves put in place, but these appear to replicate the mixed design.



explicit as to whether they fully accounted for the double effect of those control group savings. As we noted above, those control group savings are real savings for the control group and reduce the RCT estimate of treatment group savings by the same amount. That means total savings for both groups is the sum of RCT savings and two times the control group savings.

Since the report does not make an explicit issue of this either in the methods or results descriptions, it may be that the control group savings were only included once. On the one hand, this may be an object lesson of the inherent challenges of the mixed program designs. On the other, it could be seen as appropriately down weighting the less reliable opt-in program results by fifty percent. This is a serious consideration from a regulatory perspective, if commissions decide to make a distinction between the reliability of RCT and non-RCT derived behavior program savings. The downwardly biased but more reliable RCT-based savings can be claimed in full, while the additional opt-in savings are included to the extent to which they are accepted.

5.2.3.2 VIA Application

The WMECO control group savings estimates were ultimately calculated using a matched comparison group. This approach was only used after the apparent failure of each of the VIA models for the three separate waves of the program. In each case, at least two pre-months were statistically different than zero, raising questions regarding the underlying assumption of the VIA approach that sites opting into the program at different times were similar. These concerns justified moving to the comparison group approach.

Interestingly, all of the three wave models also failed to provide any statistically significant evidence of post-program savings. More problematically, the third wave model did produce highly statistically significant estimates of a positive increase in consumption correlated with the number of months since start of participation. The report explains that this counterintuitive and extreme result is caused by very different pre-period winter consumption by the last third of that wave's participants and the relatively small number of sites with longer post periods.

These issues provide evidence that the VIA approach may not be useful as reported by Harding and Hsiaw. The first of these two issues may illustrate a weakness of the VIA approach that could be overcome with the inclusion of weather terms. It would be useful to know to what extent explicitly modeling weather would address these concerns. In general, though, the extreme difference in consumption characteristics does appear to undermine the theoretical expectation that households opting into the program over time are similar. It is less clear whether the test based on pre-period consumption is sufficient to determine the suitability of these assumptions regarding underlying population. In this case, the pre-period test failures were modest compared to the problematic post-period results. Could false positive results on the pre-period test support the acceptance of questionable post-period results?

The second issue, the decreasing number of sites supporting estimates of savings further away from participation, is a longstanding challenge of consumption analysis models of this type. The unexpected monthly post-period change makes it more obvious that something is amiss but, once again, unexpected results are not a reasonable general test of the validity of results. More work is needed to understand the implications of decreasing levels of support for estimates of post-period change that, unlike more constant measure-based savings, is expected to exhibit a degree of variability.

5.2.3.3 Addressing Self-Selection

Another important highlight from the MA evaluation is the position taken on the challenge of self-selection in opt-in programs. The report has an appendix section that parallels two papers that were presented at IEPEC



2013. The report argues that self-selection is not a primary concern in the estimate of savings for these programs. The position is supported by three arguments. ODC/Navigant's interpretation of these three arguments runs counter to the position taken in this report. We present their argument and then provide our perspective on the issue.

1. The MA evaluation report of the Compact Light Pilot uses the regression corrected post-only model discussed in section 5.1.3.1. In the discussion of self-selection, the report states, "*The implication is that given a model that matches on pre-pilot energy use, with regression correction as advocated by Imbens and Wooldridge(sic) (2008) . . . , we are highly likely to generate an excellent counterfactual for participants*" (Opinion Dynamics, p. 163). This statement implies that area experts Imbens and Wooldridge would consider a matched comparison group an excellent counterfactual in the opt-in behavior program scenario. In fact, the majority of the cited Imbens and Wooldridge article is about the challenges of even producing a good comparison group based on observable characteristics let alone unobservable characteristics. The regression correction is a technique recommended to control for some of the known modeling biases of the comparison group approach not selection bias.

There is no claim, in either Imbens and Wooldridge or Imbens and Abadie that the selection correction addresses or what Imbens and Wooldridge refer to as confoundedness. In contrast, they make the following statement. "*Unlike in the setting under unconfoundedness, there is not a unified set of methods for this case. In a number of special cases there are well understood methods, but there are many cases without clear recommendations*" (p. 51. Imbens, 2009). A central finding of the Imbens and Wooldridge article is that in many cases there are no clear recommendations as to how to address self-selection or confoundedness. The regression correction approach is not presented in the article as a way to deal with confoundedness. In fact, none of the special cases discussed appear to be helpful with the challenge of self-selection in consumption modeling.

2. The report also appears to fail to identify any good reasons for thinking self-selection might exist. The report states, "*For behavioral programs, it is difficult to develop a convincing argument for selection bias given good matches based on pre-program billing history.*" The authors make this statement without ever discussing the problem of unobservable and/or time-varying characteristics. This is in contrast to the statement in the IEPEC paper that "*It is worth mentioning that, with respect to the claim that matching addresses selection bias, matching on demographic variables implies that Z (the unobservable characteristics driving selection) is invariant over time—perfect stability—and relatively highly correlated with the matched demographic variables*" (Provencher, 2013b). That is, if we are willing to assume that unobservable characteristics are both time-invariant and highly correlated with observable and available data, then selection bias should not be a problem. It is not hard to come up with counter-examples to those convenient assumptions.
3. Finally, the report provides what is referred to as a pseudo test of bias. The test compares the difference between treatment and matched comparison groups after the match period, but prior to opting-into the program. The suggestion is that selection bias would lead to diverging average consumptions and that there is no evidence of this divergence. If, however, participation is a response in part to other changes occurring at the site, it is precisely at the time of participation that we would expect the divergence to begin in the counterfactual universe that we do not get to see. Testing for that divergence in the pre-period provides no information on the possibility of self-



selection. Once again, the arguments against the importance of selection appear to avoid the specific reason selection is such a challenge.



6 CASE STUDY: PG&E PROGRESSIVE ENERGY AUDIT TOOL

The initial purpose of this project was scope out an evaluation for the PG&E Progressive Energy Audit Tool. 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 the wider horizons evident in this report. The more general question of whether opt-in behavior 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. Of equal importance, the research into opt-in behavior-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 for 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.

7 APPROACHES NOT BASED ON CONSUMPTION DATA ANALYSIS FOR OPT-IN AND OPT-OUT

7.1 Analysis of Explicit Action and Behavior Changes

An alternative general approach to examining changes in consumption is to identify individual actions and behavior changes, determine the effect of the program on those changes, and quantify the combined effects of those premise changes on energy consumption. A recent example addressing an information-only program with no tracked participants or action is in Bodman et al (2013).

Key steps with this approach include the following:

- Survey a random sample of opt-ins and opt-out customers; for the program described by Bodman et al, there was no tracking of opt-in, so this survey included the additional step of screening customers to identify those who had and hadn't engaged with the program information provided by the program.
- Determine from survey responses what actions each customer took, with and without assistance from other programs offered by the utility.
- Apply engineering estimates to determine savings from the actions
- Determine the program effect by comparing engaged/opt-in customers and unengaged/opt-out customers. A first-order estimate can be obtained by simple differences between incidence of efficiency measures or estimated savings between the two groups of customers. A more meaningful estimate is developed via regression models that isolate the program effect from underlying differences between those who opt in and those who do not.
- An additional step that can be included is to follow up with onsite inspections to confirm physical measures adopted, and obtain engineering parameters. This step can add considerable cost, and introduces additional non-response bias. To the extent the prior condition is the appropriate baseline, post-program inspections do little to resolve the key uncertainty as to what the prior condition was.

These approaches have some promise. Some evaluations have provided defensible evidence of program-induced savings using these approaches.

However, determining changes in behavior can be much more difficult than determining measure installation. Even determining measure installation outside of program tracking is subject to inaccuracy. Comprehensively identifying all relevant actions taken, determining the actions attributable to the program, and assessing the combined effect of these attributable actions is, practically speaking, impossible.

In essence, approaches that depend on identifying actions induced by the program face all the problems of self-selection and low average savings that the consumption analysis faces. These action-based approaches face the additional challenge of not having comprehensive, objective information on the actions themselves. The consumption analysis has the advantage of a reliable direct measurement (consumption) for all customers. Thus, this paper has focused on consumption data methods. Some of the issues and techniques described in this context in turn can inform evaluations that rely on analysis of actions and behaviors.



7.2 Comparison Regions

Another approach that can be considered for informational programs is comparison between a geographic area that's been exposed to such a program and a similar area without the program. This approach can be expanded to comparison across multiple regions, controlling for regional economic and demographic characteristics. The comparison can be of consumption itself, or of key behaviors targeted by the program, such as adoption of efficient equipment or changes in operating practices.

The regional approach may be effective for study of a pilot program within a service territory, where all other offerings, prices, and weather will be similar for the region where the program is and isn't offered, and comparison areas can be selected based on similar demographics and economics. This approach is most useful for community-based programs, where a concentrated local effort has the potential to produce savings large enough to be visible in a difference of differences comparison between communities.

Outside of this pilot context, regional comparisons are hard put to isolate effects of individual programs, whether behavioral or not. If the effect of interest is small and diffuse, as for a typical behavioral program, the effects of interest will not be credibly distinguishable from regional variation. Regions cannot be matched as tightly as with individual customers; and changes over time may vary across regions for many reasons.

8 CONCLUSIONS

8.1 Recognizing the Challenge

Recent behavioral programs using 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.

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 modeling, 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 behavioral 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.

8.2 Recommendations

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 CA 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 of difference approach. With this approach, three alternative results should be reported: VIA with no comparison group, rolling matched comparison group, and the combined VIA difference of difference. None of these approaches completely addresses the concern re 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 skeptically if there is a substantial portion of the participant group that has few good matches among the nonparticipants.** 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 nonparticipants 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 behavior program savings when assembling a total portfolio claim.** The joint savings are the incremental claimed savings from other programs that were induced by the behavior 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-behavioral program. The joint savings subtracted should be the incremental claim under the other program(s).

8.3 Improving Available 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 modeling approaches for opt-in behavior 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.



9 CITATIONS

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

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 Behavior and Organization*, 2013.

Opinion Dynamics. Massachusetts Three Year Cross-Cutting Behavioral Program Evaluation Integrated Report: Prepared for Massachusetts Energy Efficiency Advisory Council & Behavioral 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 Behavioral Programs." International Energy Program Evaluation Conference: Getting it Done! Evaluation Today, Better Programs Tommorrow. Chicago, IL. 2013a.

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

State and Local Energy Efficiency Action (SEEAction) Network. *Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-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.



ABOUT DNV GL

Driven by our purpose of safeguarding life, property and the environment, DNV GL enables organizations to advance the safety and sustainability of their business. We provide classification and technical assurance along with software and independent expert advisory services to the maritime, oil and gas, and energy industries. We also provide certification services to customers across a wide range of industries. Operating in more than 100 countries, our 16,000 professionals are dedicated to helping our customers make the world safer, smarter and greener.