For Trial Programs, It's Not All about Energy Savings: A Methodology for Evaluating a Retail Plug Load Program Trial

Todd E. Malinick and AJ Howard, Energy Market Innovations
Richard Ridge, Ridge & Associates
Brian Smith, Pacific Gas & Electric Company

ABSTRACT

California utilities have aggressive goals for achieving energy savings from consumer electronics and home appliances. While declines in per-unit energy consumption are evident with many product categories, the number of these "plug load" products in residences is growing, threatening the cost effectiveness of typical rebate-based programs. In response, a new generation of midstream program designs provides an opportunity to move beyond traditional "widget-based" rebate approaches by aggregating small per-unit savings across a large number of units, but these program designs are challenging to evaluate.

This paper presents the evaluation methods being used for the new Retail Plug Load Portfolio Program Trial being conducted by Pacific Gas and Electric Company and Sacramento Municipal Utilities District. The trial provides per-unit incentives to a retailer for sales of qualified products. Energy savings for the program will be measured by changes in the primary metric—sales-weighted unit energy consumption within product categories. However, given the trial's small scale and short one-year duration, detecting energy savings may be challenging. But changes in the primary metric are only one indicator of program success.

To assess the range of potential outcomes and impacts of the program, we leverage multiple methodologies to: (1) validate program theory, (2) measure quantitative and qualitative primary and proximate metrics, and (3) inform future evaluation methods for a full-scale program rollout. This paper discusses these multiple methods in detail and provides insights into how the effectiveness of future program designs can be tested using a "preponderance of evidence" approach.

Introduction

Estimates suggest that home appliances and consumer electronics—together "plug load" products—consume between 15 and 30 percent of a typical home's electrical usage (EIA 2013*a*, Navigant Consulting 2013). Today, plug loads are also the fastest growing energy use category nationwide (Kwatra, Amann, and Sachs 2013). In many home appliance and some consumer electronics product categories average unit energy consumption (UEC) is decreasing due to incremental gains in efficiency bolstered by programs like ENERGY STAR as well as general efficiency standards. However, these UEC declines are offset by net growth in the total number of products in use (Kwatra, Amann, and Sachs 2013, EIA 2013*b*). Also, though efficiency standards for many products—especially in the home appliance category—have been increasing, the fact is there are still rather inefficient models available in the marketplace. Thus, even while some products are becoming more efficient, total electricity demand from these products is still expected to increase as consumers purchase and use additional devices.

To address growing plug loads, the California Long-Term Energy Efficiency Strategic Plan (CPUC 1998) created by the California Public Utilities Commission tasked the investor-owned utilities (IOUs) of the state with reducing electricity use from plug load products by 25% by 2020. To achieve this aggressive goal, the IOUs must develop innovative approaches to program design to address plug loads consumed by consumer electronics and home appliances.

Traditional downstream programs pay rebates to consumers to prompt the purchase of more efficient products. The rationale is that because more efficient products are typically more costly, reducing the price paid by consumers by means of rebates paid directly to them will motivate them to buy more efficient products. However, since the amount of the rebate needs to be tied to the expected energy savings to make these programs cost effective, their usefulness is limited to particular product categories. Downstream, "widget-based" programs are more effective for high-consuming products like HVAC systems, because rebates can be relatively sizable. However, consumer rebates are less effective at prompting energy-efficient purchasing behaviors with products like consumer electronics and some home appliances, where energy consumption is relatively low and therefore the associated rebates are relatively small. For example, when considering the numerous factors that go into making a product purchase decision, it is unlikely that a modest $4 rebate on a Blu-ray DVD player would motivate many consumers to buy a more efficient model even while a sizable $500 rebate on an HVAC system very well may. Clearly, novel approaches and programs are needed to affect energy savings from home appliances and consumer electronics in a cost effective manner.

The remainder of this paper discusses a novel program trial currently underway aimed at curtailing plug load. The following section discusses, in detail, the theory and logic of the program trial, along with the trial objectives. However, because this is a new and novel approach aimed at affecting plug load, there are also significant challenges to evaluating the Trial and goals that go above-and-beyond just assessing the trial itself—namely assessing the array of approaches that may offer promise for evaluating a full-scale version of the program. Thus, the final sections will discuss each of the multiple methods that will be employed to evaluate the trial.

Retail Plug Load Portfolio Program Trial

The Retail Plug Load Portfolio ("RPP") Program Trial (the Trial) is being conducted to address the need for plug load programs in categories for which downstream rebate programs have been neither cost-effective nor motivational for consumers. The Trial is underway in the Pacific Gas & Electric (PG&E) and Sacramento Municipal Utility District (SMUD) service territories with a single participating major big box retailer over the October 2013 through October 2014 time period. A total of 26 retailer stores are taking part in the Trial. There is a similar trial managed by the Northwest Energy Efficiency Alliance (NEEA) underway in the Pacific Northwest.

Instead of the traditional focus on stimulating consumers to *buy* more energy-efficient products, the intent of the Trial is to generate electric energy savings by influencing the retailer to *sell* more energy-efficient models. The Trial focuses on seven targeted categories of home appliances and consumer electronics. Retailers are paid a per-unit incentive for every program-qualified model sold during the Trial. Program-qualified models are typically ENERGY STAR models in each product category, which are more efficient than comparable models within the retailer's existing product category assortment. The categories included in the RPP Trial are: (1)

room air cleaners, (2) stereos (compact audio and docking stations), (3) home-theatres-in-a-box, (4) DVD/Blu-ray players, (5) freezers, (6) refrigerators, and (7) room air conditioners.

Overall the objective is to increase the relative assortment of energy efficient products over less-efficient products. By attempting to increase sales of energy efficient models within the targeted product categories, the Trial attempts to reduce the sales-weighted unit energy consumption (SWUEC) for the targeted product categories sold by the participating retailer as a whole, thereby generating energy savings. The SWUECs are computed at the product level and are calculated as the average unit energy consumption value for all units sold by the retailer during a specific time period adjusted for the relative market share of the different models sold.

As part of the Trial, the participating retailer has committed to creating and implementing a strategic marketing plan for increasing the sales of energy-efficient models in the target product categories. Some strategies for increasing sales of energy-efficient models include: product assortment changes (e.g., stocking additional qualified models), product stocking changes (e.g., stocking more energy-efficient units), product pricing strategies (e.g., reducing prices or holding price promotions for the most efficient model in a product category), product promotion strategies (e.g., promoting energy-efficient models), product placement strategies (e.g. devoting "prime" shelf and/or store locations to energy-efficient models), and staff training (e.g., educating sales associates and managers on the benefits of energy-efficient products).

The RPP program theory is built upon a series of causal linkages between program actions and intended program outputs and outcomes. These actions, outputs, and outcomes, and the relationships between them are documented in the RPP Program Logic Model shown in Figure 1. The Trial (including the lead-up to the Trial) will encompass logic model items A-M and several of the activities, outputs, and outcomes will translate to proximate metrics (both quantitative and qualitative) that will be tracked over the life of the Trial (e.g., reduction in retailer market barriers, reduction in customer market barriers, development of marketing strategies, increases in customer purchases of program qualified products, retailer receipt of incentives, etc.).

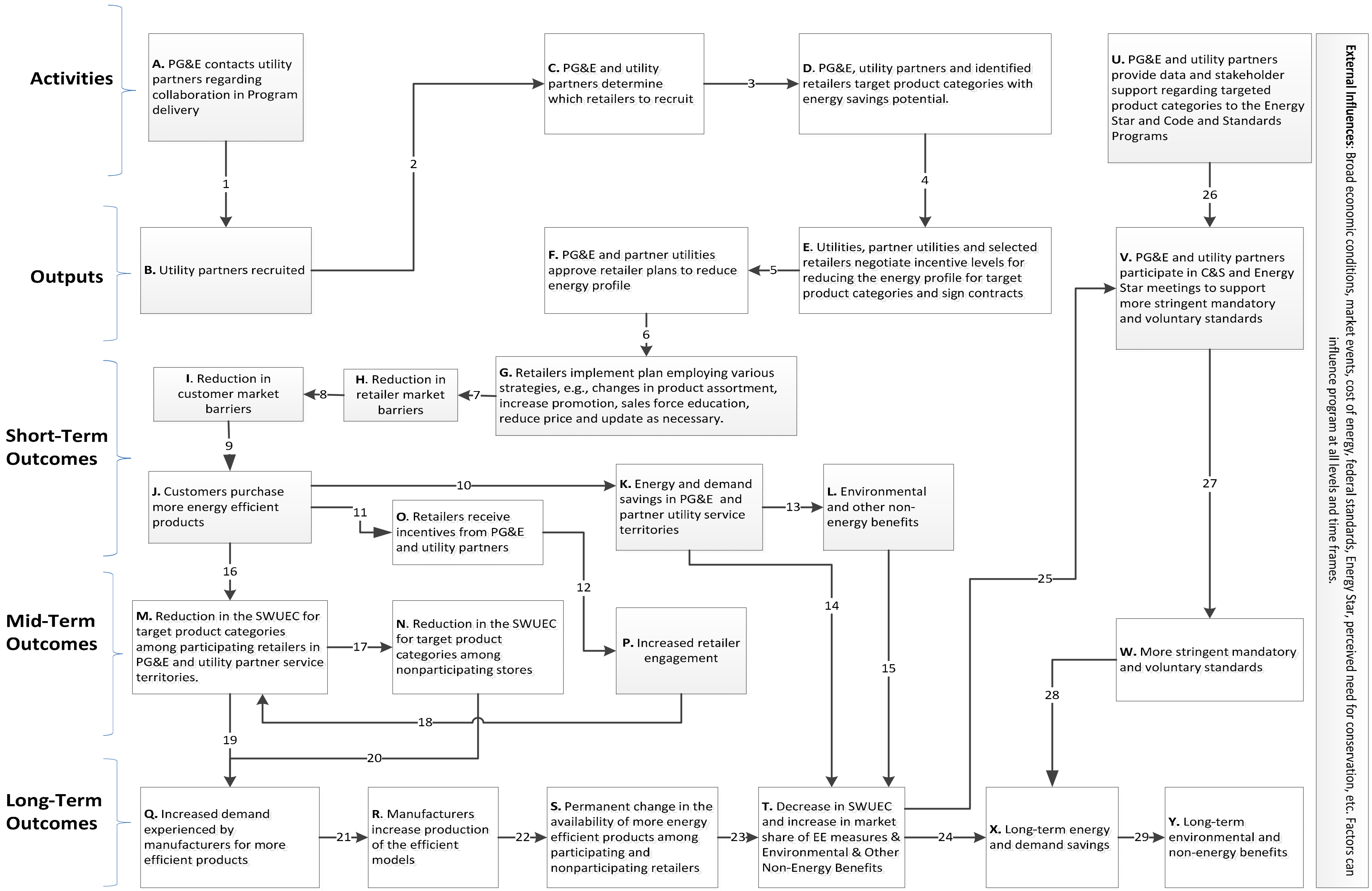

Figure 1. Retail Product Portfolio Program logic model.

RPP Program Trial Objectives

The objectives for the Trial fall into three categories: (1) performance objectives, (2) evaluation objectives, and (3) operational objectives.

Performance objectives. The performance objectives are aimed at achieving a measureable change in key performance metrics from baseline values. The performance objectives include:

- Achieve an increase in sales of program-qualified products by the participating retailer
- Achieve a reduction in the sales weighted unit energy consumption (SWUEC) of targeted product categories

Evaluation objectives. Evaluation objectives are aimed at assessing an array of factors that can point to the efficacy of the program. Also, because this is a new and novel program that poses evaluation challenges, a notable objective includes assessing the various methods that can be used to evaluate the Trial. The evaluation objectives include:

- Test various methods to evaluate the Trial
- Identify, operationalize, track, and analyze key proximate (short and mid-term) indicators of performance
- Identify data collection and analysis responsibilities (retailer staff, utility staff, evaluators)
- Measure the extent to which trial RPP objectives were achieved
- Propose performance metrics for assessing longer-term benefits
- Assess potential program trial spillover effects on non-participating stores within the PG&E and SMUD service territories

Operational objectives. The operational objectives are aimed at building a robust framework for program implementation and evaluation that could support the scaling of the RPP program concept from a trial to a full-scale, multi-year program. The following are the operational objectives that are part of the Trial:

- Retailer develops strategic marketing plan to sell more targeted products
- Retailer demonstrates implementation of the strategic marketing plan designed to sell targeted products
- Implement activities A through F denoted in the logic model (see Figure 1)
- Establish retailer data collection requirements
- Establish data processing and analysis methodology
- Identify variables that should be tracked by the RPP Program
- Determine appropriate incentive levels for a scaled-up program
- Establish a timely quality control review of retailer sales and incentive payments

- Continue building relationships with other program partners such as the Northwest Energy Efficiency Alliance (NEEA), and other California utilities to increase the proportion of U.S. households addressed by a larger program
- Increase retailer awareness of promoting, selling, and/or stocking products based on energy efficiency criteria

Taking Stock: Looking Ahead—Considerations for Evaluating the RPP Program Trial

The evaluation approach chosen for the RPP Program Trial is a multifaceted approach consisting of several different tactics (see next section). Multiple tactics were chosen because this trial is a relatively new and novel program design and in addition to taking stock of the current trial itself, a significant goal is to also look ahead and assess what approaches may prove useful for evaluating a full-scale rollout of such a program. This section of the paper discusses these issues and discusses some other considerations regarding the development of the evaluation approach.

Because of the diversity of performance, evaluation, and operational objectives that exist for the Trial, no single evaluation methodology is capable of assessing all objectives. Some objectives are more qualitative in nature and involve assessing and evaluating operational activities and processes to ensure that the program trial is being implemented as planned and functioning as expected. Other objectives are more quantitative in nature and involve defining, measuring, and analyzing specific metrics that will serve as indicators of program trial progress and/or success. Thus, multiple, industry-accepted evaluation methodologies are being employed.

However, it is also important to emphasize that this is a *trial* of a relatively new and novel program concept. Because of this, there are additional objectives that go beyond just evaluating the trial itself. Instead, some of the additional goals are aimed at: (1) assessing various potential evaluation approaches to gauge which tactics may prove useful for evaluating a full-scale roll-out of the program, and (2) assessing what information is needed and available under each of the evaluation approaches. The evaluation team (the authors of this paper) did not take part in the actual design of the RPP Program Trial, but have provided input during the Trial aimed at improving implementation and ensuring the evaluability of the Trial and a potential program.

Since a primary performance objective of the Trial is to affect net reductions in energy consumption for targeted product categories, several of the evaluation approaches are focused on evaluating changes in the primary metric—SWUEC. The evaluation team selected several approaches ranging in sophistication—and in the data needed to conduct the modeling—to assess which, if any of the approaches might prove useful for detecting the energy impacts of the program. These approaches include quasi-experimental non-equivalent participant/non-participant comparison group-based designs (both within service territory and statewide) and participant-only designs (i.e. quasi-experimental pre-post segmented regression and quasi-experimental forecasted baseline to recorded data). However, it should be emphasized that changes in the SWUEC are mid-term outcomes and are unlikely to be substantial during the relatively short timeframe of the trial, because for some of the products, UECs are quite low to begin with and the incremental differences in UEC associated with more energy-efficient models are quite small.

Changes in the SWUEC will be a function of many factors such as what activities the retailers ultimately implement, when they implement the activities, and to what scale or extent. While, in principle, large enough sample sizes could allow for detecting small effect sizes, the small scale and short timeframe of this Trial (i.e. a single retailer in only 26 participating stores for about a year), is not likely enough. That said, it is important to reiterate that a notable goal of this study is to assess the potential value of the different approaches for evaluating a full-scale version of the program. Thus, even though statistically significant differences may not be evident as a result of the current trial, this is a critical component of the evaluation approach because some manner of estimating energy savings with a reasonable degree of accuracy, while also being able to attribute the savings to activities undertaken as part of the program, will likely be demanded by regulators. Importantly, incorporating these efforts into the evaluation of the Trial will give the evaluation team the ability to make recommendations to program staff and regulators into the general evaluability of any future program.

In terms of the current Trial, the threat is that total reliance on quantitative analyses of the changes to SWUECs will fail to assess the influence of the Trial on retailer behavior and inadequately reflect the overall promise of new and novel programs aimed at reducing plug load. The Trial is primarily a program aimed at transforming the market, and market transformation through intervention by utilities usually takes years to occur. Thus, other factors other than SWUEC need to be taken into consideration when devising an evaluation approach that can point to and be indicative of the longer-term impacts and effectiveness of this Trial. Thus, our evaluation approach includes additional methods and tasks aimed at making a confirmatory assessment of program efficacy and verification of key causal mechanisms and outcomes. Also, because this is a trial, these impact evaluation efforts will also yield important information regarding the implementation of the program trial and provide early feedback so that mid-course corrections can be made if necessary. That said, to the extent possible, the evaluation of the Trial will comply with the *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals* (TecMarket Works Team 2005).

Evaluation Methods

In order to fully capture potential impacts of the Trial the evaluation team is using a mixed-methods evaluation approach. This approach will utilize the concept of triangulation, or the attempt to measure something using several independent routes. Three key benefits of this triangulation approach are that: (1) it is not dependent on the validity of any single source, (2) it allows us to assess a very wide range of research methods, and (3) it will allow the evaluation team to understand which of the proposed methods are best suited to the RPP program concept.

It is expected that some of these methods will be used to evaluate a full-scale program implementation, others may ultimately only be appropriate for evaluating the Trial, and others may prove totally ineffective for either. Also, the list of methods should not be considered comprehensive, as additional approaches will likely emerge to capture all aspects of a full-scale implementation.

Figure 2 shows the six research designs being used. The modeling approaches (in blue in Figure 2) were selected because the evaluation team, program team, and consulting representatives of the California Public Utilities Commission – Energy Division feel they have promise for quantifying energy savings and attribution of these savings to the program. The more qualitative approaches (in orange in Figure 2) were selected because they can get at the other, more procedural aspects of the program implementation. None of the approaches are new to energy efficiency program evaluation; it is the combination of the multiple methods that is novel. Each of these approaches is discussed in the subsequent sections.

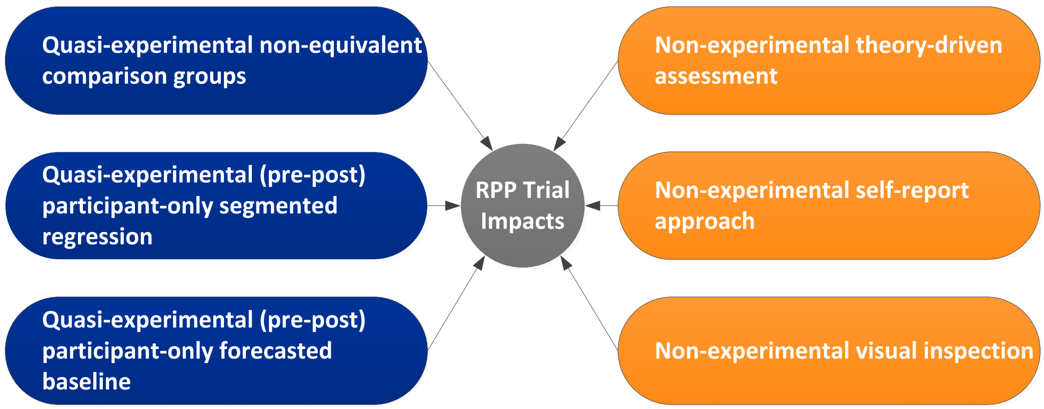

Figure 2. RPP Program Trial Research Designs.

Quasi-Experimental Non-Equivalent Comparison Group Designs

When unable to implement a true controlled experimental design in which group differences are controlled for through random assignment, there is a wide variety of quasi-experimental designs available. One such design, the non-equivalent comparison group design, does not involve random assignment to treatment and control groups, and, as a result, the two groups are not considered strictly equivalent in all aspects other than the intervention. Instead, this approach relies on defining treatment and controls based on some inherent grouping in the general population. This design adequately controls for many of the threats to internal validity, except for self-selection. A *self-selection bias* would result when there are systematic differences between the stores included in the trial and those that were excluded from it in a manner that affects the SWUEC. A matching process will be used to facilitate comparisons of stores that are similar across multiple characteristics including store size, sales volume, and climate zone. In order to conduct the most valid possible comparisons and reduce the threat of self-selection bias, the evaluation team will initiate a matching process to ensure appropriate non-participating stores are paired with and compared against similar participating stores.

The evaluation team will conduct two non-equivalent comparison group analyses: (1) participant stores to non-participant stores in the PG&E/SMUD service territories, and (2) participant stores to non-participant stores in the State of California. For each of these comparisons, the evaluation team will utilize segmented regression analysis with a non-equivalent control group, which is a powerful statistical method for estimating intervention effects in interrupted time series studies (Wagner et al. 2002), to model and assess changes in SWUECs. With this approach, the goal is to determine if a statistically significant difference exists between the modeled SWUECs between participating and non-participating stores. Figure 3 illustrates this method.

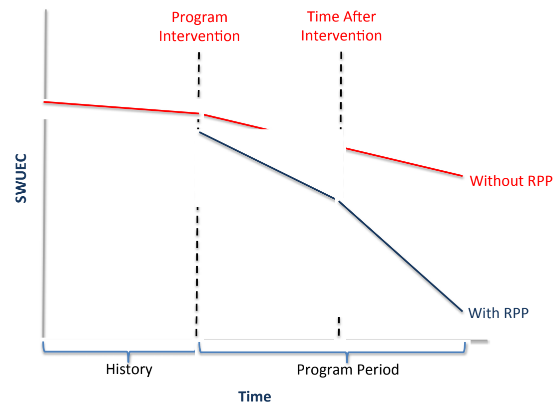

Figure 3. Illustration of segmented regression
Approach with a non-equivalent control group.

Note from Figure 3 that three time periods are defined: (1) a pre-program, historical period, (2) a period in time where a specific intervention tactic occurs (e.g., a particular promotion, a price reduction, alterations to stocking assortment, etc.), and (3) a period of time after a specific intervention tactic, but prior to the termination of the Trial. In the hypothetical example shown, the historical SWUEC is plotted up till the beginning of the intervention (in green), then there are two plots: one for the non-participating control stores (in red labeled "without RPP") and one for the participating stores (in blue labeled "with RPP"). The effect of the program would be the difference in SWUEC between the "without RPP" and "with RPP" segments. Another aspect of the Trial that Figure 3 highlights is that the slope of the "with RPP" line can increase in slope after a particular tactic ceases if there are lagged effects. For example, a certain tactic may lead to increased sales of the promoted product, which in turn can lead to the retailer to decide to alter assortments of the product category, maybe adding additional program qualified models. This is exactly the type of effect expected and included in the program theory and logic model.

For this quasi-experimental design and the two that follow, there is an additional concern: given the one year duration of this trial, the participating retailer might not have enough time to fully implement the various strategies (e.g., product location, product placement, advertising, and assortment) designed to increase sales of program-qualified products. As a result, substantial shifts in the SWUEC are unlikely. This threat increases the importance of other, proximate indicators tied to the program theory activities, outputs, and outcomes as outlined in the logic model (see Figure 1), which are the focus of the non-experimental designs. If proximate indicators are moving in the right direction, then our confidence is increased that substantial changes in the SWUEC will eventually occur over the course of a longer program period.

Quasi-Experimental Pre-Post Participants-Only Design Using Segmented Regression

A second approach we will use to assess the influence of the program on SWUEC is a quasi-experimental pre-post participants-only design using segmented regression. This approach uses the participating stores as their own control and analyzes the pre- and post-intervention measurements of the SWUEC and attempts to measure the change in the slopes and or intercepts from pre-period to the post period. Figure 4 illustrates this approach.

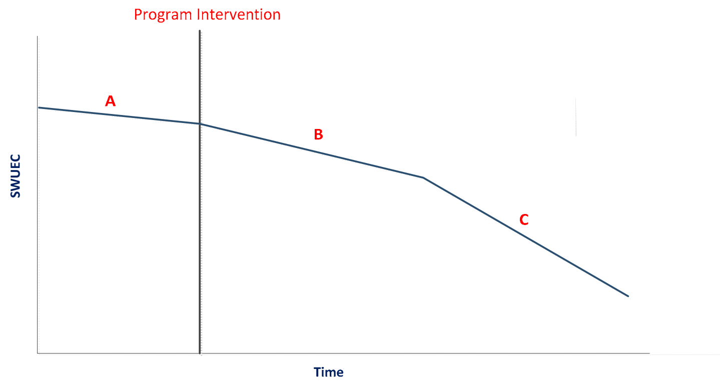

Figure 4. Illustration of segmented regression approach with treatment group only.

As shown in Figure 4, the evaluation team will attempt to determine whether Slopes B and/or C differ from the baseline Slope A. For example, depending on the participating retailers implementation strategy, Slope B may be different than Slope A due to initial product placement and in-store advertising; Slope C might be different than Slopes A or B due to later changes in assortment.

History, or the possibility that some other event besides the program-related interventions might be causing some or most of the observed effects, is a threat to the internal validity of this approach. However, the evaluation team will be using monthly data, which can help to mitigate this threat since the historical events that can explain an apparent treatment effect are fewer with monthly data than with yearly (Shadish, Cook, and Campbell 2002). That is, the more granular the data, the more clearly that specific events can be tied to effects in particular months. In addition, a list of other plausible effect-causing events (such as changes in codes and standards or natural technological advancement) that could influence the SWUEC could be identified and, using quantitative or qualitative means, the evaluation team can seek to determine whether most or all of them appear to have affected any changes in slope.

Quasi-Experimental Pre-Post Design Comparing the Forecasted Participant Baseline to Recorded Data

A third quasi-experimental design that will be used to assess program effects involves using the available historical sales data from participant stores to forecast a counterfactual baseline over a 12-month horizon, which is then compared to the actual participant sales data post-program implementation for the same time period. The difference between the recorded and the forecasted baseline is the net effect of the program. Figure 5 illustrates this approach.

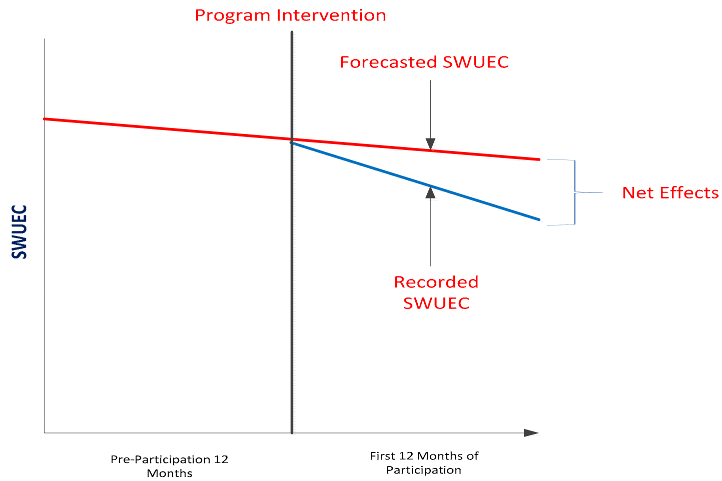

Figure 5. Forecasted monthly SWUEC versus recorded.

Similar to the discussion around the quasi-experimental participant-only design using segmented regression, the main threat to internal validity for this approach is history. However, like the segmented regression approach, the use of monthly data can help to mitigate this threat and the same list of other plausible effect-causing events that could influence the SWUEC can be used to help determine whether any of these influences appear related to any net differences.

However, for this approach, there are two additional concerns. First, because only 20 months of historical data will be available and there is, in some cases, substantial variability in the historical data, the precision of the forecasts will be affected (i.e., the confidence interval around any one forecast might be rather large and include the recorded SWUEC). This matters because unlike the other approaches, this approach is so highly dependent on the historical data to determine net effects. Second, the adoption during the Trial of any new efficiency standards for targeted products will introduce a potentially serious confounding variable (i.e., one that competes with the Trial in explaining changes in the SWUEC). The evaluation team will carefully assess these issues when conducting and interpreting our analyses.

Non-Experimental Theory-Driven Assessment

One of the stronger non-experimental methods, developed by a number of evaluators is the theory-driven evaluation (Weiss 1997, 1998). This approach is often implemented by undertaking three steps: (1) developing program impact theory, (2) formulating and prioritizing evaluation questions, and (3) answering evaluation questions (Donaldson 2007).

This approach involves operationalizing key performance indicators associated with key causal linkages (i.e., converting the performance indicators into quantifiable and measureable metrics) in the logic model. If predicted steps between an activity and an outcome can be confirmed in implementation, this matching of the theory to observed outcomes will lend a strong argument for causality (Weiss 1997).

The evaluation team and utility staff have worked together to develop a robust logic model outlining the activities, outputs, and short-, mid- and long-term outcomes associated with the program (see Figure 1). Through the use of proximate metrics tied to these activities, outputs, and short-term outcomes, the theory-driven assessment will be used to validate the program theory and logic to ensure the overall program design is working as intended.

A theory-driven approach is particularly useful when a new approach is being undertaken since many of the cause-and-effect relationships might be relatively untested and implementation problems more numerous. This is the case with the Trial, which deviates from the traditional program design and embraces a mid-market intervention that focuses on a portfolio of efficient products rather than a single widget-based approach.

It is worth noting that, though a number of components of the logic model can be tested directly through the evaluation of this trial, not all components can be tested directly. For example, some of the hypothesized outcomes are mid- to longer-term, and are not expected to be observable during the short one-year duration of the Trial.

Validation of key activities, outputs, and outcomes as well as the hypothesized causal linkages among them as illustrated in the logic model and supported in the program theory is one of the major objectives of this study. Confirming the key linkages within this theory-driven framework will provide an alternative path to assessing program efficacy.

Self-Report Approach

In 1993, the recognition that methods involving comparison groups were not always feasible was first formalized in the energy efficiency field in California in the *Protocols and Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management Programs* (CPUC 1998). For budgetary, timing, statistical, and research design reasons, more traditional designs and analyses must sometimes be replaced or supplemented with the self-report approach. The self-report approach can include both quantitative and qualitative information and can consist of data collection efforts such as surveys and in-depth interviews, as well as information such as monthly implementation updates provided by the retailer.

For the Trial, various market actors can report their likely behaviors absent the RPP intervention, the influence of the Trial on their behaviors, as well as the extent to which the program trial was successfully implemented. For example, interviews with key actors such as participating retailer product buyers can provide information related to their decisions to buy and change the assortment of targeted products. Marketers can provide information related to their decisions to promote and uniquely position targeted products within the stores. Utility and implementation staff can provide insights into how well the program is operating, what worked, what did not, concerns, and recommendations for improving the program moving forward.

Many of the evaluation objectives for this trial can only be answered via information gathered directly from market actors. This is especially the case with the topics that are more operational in nature. For example, the evaluation team will need to talk with retailer and utility staff to understand what they are doing, how they are doing it, why they are doing it, and what they might change to improve program implementation. However, the self-report approach can also be used to support the theory-driven assessment by soliciting input from key stakeholders on issues such as the examples mentioned above in the “considerations” section.

Visual Inspection

While quantitative analysis can be compelling, one should not discount the value of visual inspection of the SWUEC trends. Though wide confidence bands and uncertainty over the influence of future code changes will likely make the detection of program-induced differences difficult to detect from a strict statistical perspective, from a practical perspective the graphical representations will be quite helpful for detecting early trends that may not be significant now, but visual evidence might suggest could become statistically significant in the future. By visual inspection, one often obtains an immediate and strong impression of the trend just by noting whether the series seems flat or moves generally up or down over time. The transfer function(s) (the empirical change in the times series after the intervention point) is assessed by using a vertical line at the intervention point(s) and looking separately at the before and after series—much like the examples shown in Figures 3, 4 and 5. One notes whether the same trend has carried over from the before to the after and whether there seems to be a change in the overall level of the series that occurs at the intervention point or a bit later.

Though visual inspection can be fallible in the face of small or delayed effects (Furlong and Wampold 1981; Ottenbacher 1986), this approach is still useful given that it has the ability to reveal patterns that are apparent to the eye, but not detectable through more rigorous statistical methods. A threat to this approach is that an apparent trend found through this approach may not actually increase or persist into the future. However, this approach—like the other approaches discussed herein—will not be used in isolation, and will instead be interpreted in relation to other findings to paint the bigger picture and help to ensure errant conclusions are not drawn.

Our confidence in the results of the more rigorous analyses described in this plan will be increased to the extent that they are supported by visual inspection. Also, though statistical significance may not be reached in some analyses for this trial with any reasonable level of significance, visual inspection can reveal if non-significant results are at least trending in the hypothesized direction, thus providing some support for the notion that they could become significant were the program to be implemented at a larger scale and/or for a longer duration. Any conclusions drawn from this approach that are not consistent with the more rigorous methods will be carefully qualified when reporting results.

Conclusions

Interventions that intend to promote the sales of plug-load measures whose per-unit savings and rebates are relatively small must deviate from the traditional downstream rebate program design. Such programs may be cost-effectively targeted at midstream actors such as retailers, who can focus their efforts on promoting *portfolios* of products rather than *single* products. Of course, since evaluation protocols used in many jurisdictions were designed for more traditional programs designs, innovative evaluation methods are needed to assess the performance of these novel interventions. These new methods must be assessed before they can be recommended to evaluators for widespread use. Multiple methods should be used for innovative programs given the inherent uncertainty around the results from any one method. In parallel with these efforts, regulatory bodies must assess what will be deemed as credible evidence in light of the constraints placed on utilities that are being compelled to test innovative program designs that do not lend themselves to evaluation by any single evaluation method.

References

Box, G. E. P. and G. M. Jenkins. 1970. *Time-Series Analysis: Forecasting and Control*. San Francisco: Holden Day.

CPUC (California Public Utilities Commission). 2008. *California Long-Term Energy Efficiency Strategic Plan.* http://www.cpuc.ca.gov/NR/rdonlyres/D4321448-208C-48F9-9F62-1BBB14A8D717/0/EEStrategicPlan.pdf

CPUC (California Public Utilities Commission). 1998. *Protocols and Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management Programs.* California Public Utilities Commission Decision 93-05-063.

Donaldson, S. I. 2007. *Program Theory-Driven Evaluation Science*. New York: Lawrence Erlbaum.

EIA (Energy Information Administration). 2013*a*. *Estimated U.S. Residential Electricity Consumption by End-use, 2011.* http://www.eia.gov/tools/faqs/faq.cfm?id=96&t=3. Updated June 24, 2013.

———. 2013*b*. *Analysis and Representation of Miscellaneous Electric Loads in NEMS.* Washington, D.C.: U.S. Department of Energy. http://www.eia.gov/analysis/studies/demand/miscelectric/pdf/miscelectric.pdf

Furlong, M. J. and B. E. Wampold. 1981. "Visual Analysis of Single-Subject Studies by School Psychologists." *Psychology in Schools* 18: 80-86.

Kwatra S., J. Amann and H. Sachs. 2013. *Miscellaneous Energy Loads in Buildings*. Washington, D.C.: American Council for an Energy-Efficient Economy. http://www.aceee.org/sites/default/files/publications/researchreports/a133.pdf

Navigant Consulting, Inc. 2013. *The 2013 California Energy Efficiency Potential and Goals Study: Final Draft Report*. Prepared for the California Public Utilities Commission.

Ottenbacher, K. J. 1986. "Reliability and Accuracy of Visually Analyzing Graphed Data from Single-Subject Designs." *American Journal of Occupational Therapy* 40: 464-469.

Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin Company.

TecMarket Works Team. 2005. *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*. San Francisco: California Public Utilities Commission. http://www.calmac.org/events/EvaluatorsProtocols_Final_AdoptedviaRuling_06-19-2006.pdf

Veney, J.E. and Luckey, J. W. 1982. "A Comparison of Regression and ARIMA Models for Assessing Program Effects: An Application to the Mandated Highway Speed Limit Reduction of 1974." *Social Indicators Research* 12(1): 83-105.

Wagner et al. 2002. "Segmented Regression Analysis of Interrupted Time Series Studies in Medication Research." *Journal of Clinical Pharmacy and Therapeutics* 27: 299-309.

Weiss, C. H. 1997. "Theory-Based Evaluation: Past, Present, and Future." In D. J. Rog and D. Fournier (eds.) *Progress and Future Directions in Evaluation: Perspectives on Theory, Practice, and Methods*. San Francisco: Jossey-Bass Publishers.

———. 1998. *Evaluation*. Upper Saddle River, NJ: Prentice Hall.