Spillover as a Cause of Bias in Baseline Evaluation Methods for Demand Response Programs

Authors:

Annika Todd, Peter A. Cappers, C. Anna Spurlock, Ling Jin

Energy Analysis and Environmental Impacts Division
Lawrence Berkeley National Laboratory

September 2019

This is a pre-print copy of a journal article published in *Applied Energy*.
DOI: https://doi.org/10.1016/j.apenergy.2019.05.050.

This work was supported by the U.S. Department of Energy's Office of Electricity Delivery and Energy Reliability under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California.

Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.

COPYRIGHT NOTICE

This manuscript has been authored by an author at Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 with the U.S. Department of Energy. The U.S. Government retains, and the publisher, by accepting the article for publication, acknowledges, that the U.S. Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes.

Spillover as a Cause of Bias in Baseline Evaluation Methods for Demand Response Programs

Authors: Annika Todd[a] (Lawrence Berkeley National Laboratory), Peter Cappers[b], C. Anna Spurlock[a], Ling Jin[a*]

[a] Lawrence Berkeley National Laboratory, 1 Cyclotron Rd., Berkeley, CA 94720, USA

[b] Lawrence Berkeley National Laboratory, c/o 7847 Karakul Lane, Fayetteville, NY 13066, USA

* Corresponding author at: 1 Cyclotron Rd., Berkeley, CA 94720, USA, Tel +1 (510) 861-5148.

Email address: ljin@lbl.gov (L. Jin)

Abstract

Prior research has shown peak load reduction estimates from residential event-driven demand response programs (e.g., Critical Peak Pricing) using X of the highest Y days with a weather adjustment method are the best performing within the class of currently used baseline methods. However, they are still biased relative to estimates produced from randomized control trials (RCTs), the unbiased “gold standard” evaluation method. In this paper we identify underlying factors that cause some of the bias found in one commonly used baseline method. Rather than simply quantifying bias, this deeper understanding can be used to develop more accurate methods that are not subject to these underlying factors of bias. Previous studies have compared various baseline methods relative to each other; however, because all baseline methods are biased, it is impossible to determine the true bias that exists in them. We have access to a unique RCT dataset: the Sacramento Municipal Utility District’s (SMUD) consumer behavior study on critical peak pricing. Our analysis of the 23 event days over two summers allows us to identify the true bias on peak load reduction estimates by using the RCT estimates as the unbiased gold standard against which we compare the estimates from the baseline methods. We found that spillover of energy reductions, from hours targeted by a program onto other hours, is one underlying factor that is a major cause of bias in baseline methods. We discuss alternative baseline methods that may not be subject to this same bias.

Keywords: Baseline prediction; Demand response; Critical peak pricing; Peak demand savings; Experimental design; randomized control trials

1. Introduction

The implementation of demand response (DR) opportunities (i.e., time-based rates and incentive-based programs) for residential and small commercial electricity customers is rapidly expanding due in large part to recent broad-based deployment of Advanced Metering Infrastructure (AMI) [1]. With the ability to now more granularly measure and/or control residential customers' load and peak demand, states are beginning to commit to a broader array and greater penetration of time-based rates and incentive-based programs.

Many of these DR opportunities are event-driven. For example, Critical Peak Pricing (CPP) programs charge higher rates during event periods on a limited number of days per year in order to induce peak demand reductions [2, 3]. Therefore the need to produce accurate and unbiased estimates of the peak load reductions during events caused by these programs will be increasingly important for program cost-effectiveness evaluations, dependability and reliability assessments for resource adequacy credit, load and/or peak demand forecasting, and customer settlement.

Currently, a number of different approaches exist to estimate the peak load reduction provided by time-based rates or incentive-based programs that are event-driven, the most widely used are baseline methods [4, 5]. They generally differ along the following lines:

- Estimation methods (e.g., average, matching and regression);
- Timeframes (e.g., from same/previous day to previous year);
- Data selection rules (e.g., proximity to event, similarity of load, similarity of weather, highest or middle x of y); and
- Sensitivity to external factors (e.g., heat, humidity).

Many incentive-based programs that are directly marketed to residential and small commercial customers use a few days of usage prior to an event to create a simple baseline that is easy to describe to potential participants. At the other end of the spectrum of baseline methods, a few ISO/RTOs employ very sophisticated statistical analysis techniques which tend to attract much larger customers (e.g., industrial process plants).

Considering that these various baseline methods are all employed to ultimately quantify participants' contribution to resource adequacy or some other bulk power system service as well as for performance payment purposes, their accuracy is of critical importance.

A number of analyses have been previously undertaken to examine the accuracy of different baseline methods for estimating peak load reductions [4-11]. The most comprehensive of these compared the consistency and accuracy of several commonly applied baseline methods using a number of different metrics applied against interval meter data of a large sample of commercial and industrial participants and non-participants in the PJM footprint [4] and alternatively of residential participants in San Diego Gas & Electric's critical peak rebate (CPR) program [5]. One key result from both of these analyses was that weather can be a significant source of bias in baseline methods. As such, baseline methods that included an adjustment for weather on the day of the event performed better than methods that did not. Among the class of commonly applied baseline methods, the ones found to be the best performing averaged the load for customers across the highest 3 or 4 out of the prior 5 non-event weekdays and included an adjustment for weather on the event day.

These previous studies compared various baseline methods relative to each other. Subsequent research compared the best performing 4-in-5 day baseline method with weather adjustments (which we will refer to hereafter as the "4-in-5" method) to an entirely different class of methods for determining estimates of load impacts: randomized controlled trials (RCTs), the "gold standard" evaluation method [10].[1] Correctly designed and implemented RCTs result in a completely unbiased estimate of load reduction. By leveraging the power of experimental design employed in a U.S. Department of Energy (DOE) Smart Grid Investment Grant (SGIG) funded consumer behavior study of time-based rates, peak load reduction estimates using the 4-in-5 method were found to be significantly different (i.e., biased) relative to estimates produced from RCTs. However, that research did not identify the causes of this bias.

This paper presents an expansion of this prior research [10] by examining the underlying factors that cause some of the bias found in the 4-in-5 baseline method, in order to identify alternative methods that may not be subject to this same bias (previous studies simply quantified the bias). To this end, our research seeks to address three main questions:

1. Can we identify the **magnitude of the bias** in the 4-in-5 method?
2. Beyond weather, can we identify additional **factors that cause bias** in the 4-in-5 method?
3. Are there alternative baseline methods that are designed to **avoid factors that cause bias**?

To answer these questions, we analyze data collected as part of the SGIG consumer behavior studies of time-based rates. Specifically, this analysis examines the above research questions in the context of a CPP program with higher prices during peak hours on 23 event days over two summers in the Sacramento Municipal Utility District's (SMUD) consumer behavior study, which utilized a randomized control trial experimental design [8].[2] We use peak load reduction estimates from the RCT as the unbiased gold standard against which we compare the estimates from the 4-in-5 method as well as alternative methods.

2. Methodology

2.1. Definition of the 4-in-5 Baseline Method

The 4-in-5 baseline method is defined as follows; for each participating customer and each event:[3]

- Take the hourly interval meter data for the most recent five calendar days preceding an event, excluding weekends, holidays, days where a prior event was called, and those with other exclusion criteria;
- Identify the four pre-event days with the highest average daily electricity usage;

[1] Baylis et al. [10] built upon the pioneering work by LaLonde [12] comparing RCTs to other program evaluation methods.

[2] Clearly, CPP does not use a baseline for customer settlement like is applied in CPR programs. However, any event-driven DR opportunity, either rate or program, can be used to derive an estimate of the load reductions during events via common baseline methods used in incentive-based programs like CPR.

[3] For more details on the 4-in-5 baseline method, see **Error! Reference source not found.**B.

- Define three “baseline calculation periods,” which are the average hourly electricity usage (kWh/h) during each of the following periods: (1) pre-peak hours on the event day; (2) peak hours on the four pre-event days; and (3) pre-peak hours on the four pre-event days;
- Take an average of the hourly electricity usage during the peak hours on the four pre-event days to create an unadjusted baseline;
- To account for the fact that there may be higher overall usage during event days, measure the difference in usage between pre-peak hours on the event day relative to the pre-peak hours on the four pre-event days, and add this to the unadjusted baseline in order to create an adjusted baseline.

This definition leads to the specification that follows in Eq. 1 and Eq. 2 for estimating the peak load reduction generated by the 4-in-5 baseline method. In these equations, the input variables represent average hourly usage during certain times. *EventPeak* is usage during the hours targeted for load reduction – peak hours on the event day. The other three variables are during the three baseline calculation periods: *EventPrePeak* is usage during pre-peak hours on the event day; *BaselinePeak* is usage during peak hours on the four pre-event baseline days; and *BaselinePrePeak* is usage during pre-peak hours on the four pre-event baseline days. For each customer *i* and event *e*:

$$Baseline_{ie} = BaselinePeak_{ie} + (EventPrePeak_{ie} - BaselinePrePeak_{ie}) \qquad \text{(Eq. 1)}$$

$$Estimated\ Peak\ Load\ Reduction_{ie} = \ EventPeak_{ie} - Baseline_{ie} \qquad \text{(Eq. 2)}$$

2.2. Method for Calculating Total Bias

As previously noted, prior research tested the accuracy of this particular baseline method when applied to customers who were exposed to a critical peak pricing rate implemented as part of a randomized controlled trial [10]. First an estimate of the total program impact on peak load reduction from the 4-in-5 method was derived for each event by averaging the customer level estimate across all participants. Then the bias for each event was calculated by comparing those impact estimates from the 4-in-5 method against those produced by the RCT.

Our research builds directly off of this prior research [10]. We examine two CPP rate opportunities implemented as RCTs: one used a default enrollment approach (i.e., customers are automatically put on the rate but allowed to opt-out); and one used a voluntary enrollment approach (i.e., customers must opt-in to the rate). We use peak load reduction estimates from the RCT as the unbiased gold standard against which we compare the estimates from the 4-in-5 method to calculate bias for each event. We will hereafter refer to this as the “total bias” for each event because we later identify a specific type of bias that may contribute to the total bias. Note that we are comparing the total or specific bias for each event on average across all participants; it is impossible to develop an unbiased estimate of the bias for each individual participant using an RCT as designed in SMUD’s consumer behavior study.

2.3. Weather and Spillover Biases and their Effects on the 4-in-5 Method

One well known and intuitive source of bias in baseline methods is weather [4, 6, 7]: if events are called on the hottest days, a baseline built on the days prior to the event will likely have lower usage and therefore will underestimate the energy reduction on the event day. An adjustment can be applied to a baseline to help improve accuracy and thus remove some of the total bias associated with baselines that do not have a weather adjustment [4, 6, 7]. However, as shown in [10], even with such weather adjustments applied to the 4-in-5 baseline, other forms of bias must still exist.

In this paper, our hypothesis is that in addition to the bias caused by weather, spillover drives some of the total bias. In this context, spillover refers to the phenomenon that occurs when a program is designed to target energy reduction during specific hours or for specific energy behaviors, but customers respond to the program during other hours or for other energy behaviors in addition to those targeted by the program.[4] For time-based rates and incentive-based programs, this means that while the program may cause a customer to reduce their usage during the target hours, it also may cause the customer to reduce usage during other non-targeted hours. Specifically, in the context of using the 4-in-5 baseline method to calculate peak load reductions during CPP events, customers may reduce their usage not only during the targeted peak hours on event days, but also during the three baseline calculation periods defined in Section 2.1: (1) pre-peak hours on the event day; (2) peak hours on the four pre-event days; and (3) pre-peak hours on the four pre-event days.

If spillover exists, it has implications for the accuracy of baseline methods that rely on usage during time periods affected by spillover (see Figure 1). For the 4-in-5 baseline method, spillover onto any of the three baseline calculation periods would cause a systematic bias[5] (i.e., the bias is consistently positive or negative); specifically, it would cause the 4-in-5 baseline method to underestimate the peak load estimates during the majority of events. To see why, consider that the true baseline represents how much energy a customer would use in the *absence* of the program.[6] The 4-in-5 method assumes that the program only affects usage during the targeted peak hours, and creates the baseline with usage during other hours (the three baseline calculation periods). When spillover occurs, program participants lower usage not only during the targeted peak hours, but also during any of the three baseline calculation periods. This in turn would cause the 4-in-5 baseline to be lower than the true baseline. Then, when the peak load reduction is estimated by comparing the baseline to the actual usage during the peak hours on the event day, the estimated reduction would appear to be less than it actually is; that is, the 4-in-5 method would underestimate the true peak load reduction.

[4] Another meaning of spillover is reduction in usage outside of the *participant population* in the CPP rate (e.g., customers on the CPP rate talk to their neighbors about the importance of reducing usage during peak hours who then do so without the economic incentives of being on the CPP rate).

[5] Bias measures the directional difference between the expected value of the estimates and the truth, while precision measures variations of estimates within themselves. We use overall accuracy to encompass both bias and precision (e.g., the root mean squared error (RMSE) as a measure of accuracy incorporates both deviation from the truth and variance).

[6] In this paper, we focus on evaluation of the impact of a program on peak load reduction relative to the counterfactual in which the program is absent. Note that this is different from a day- or week-ahead forecast of the impact of calling an event on one specific day *given* that the program is already in place.

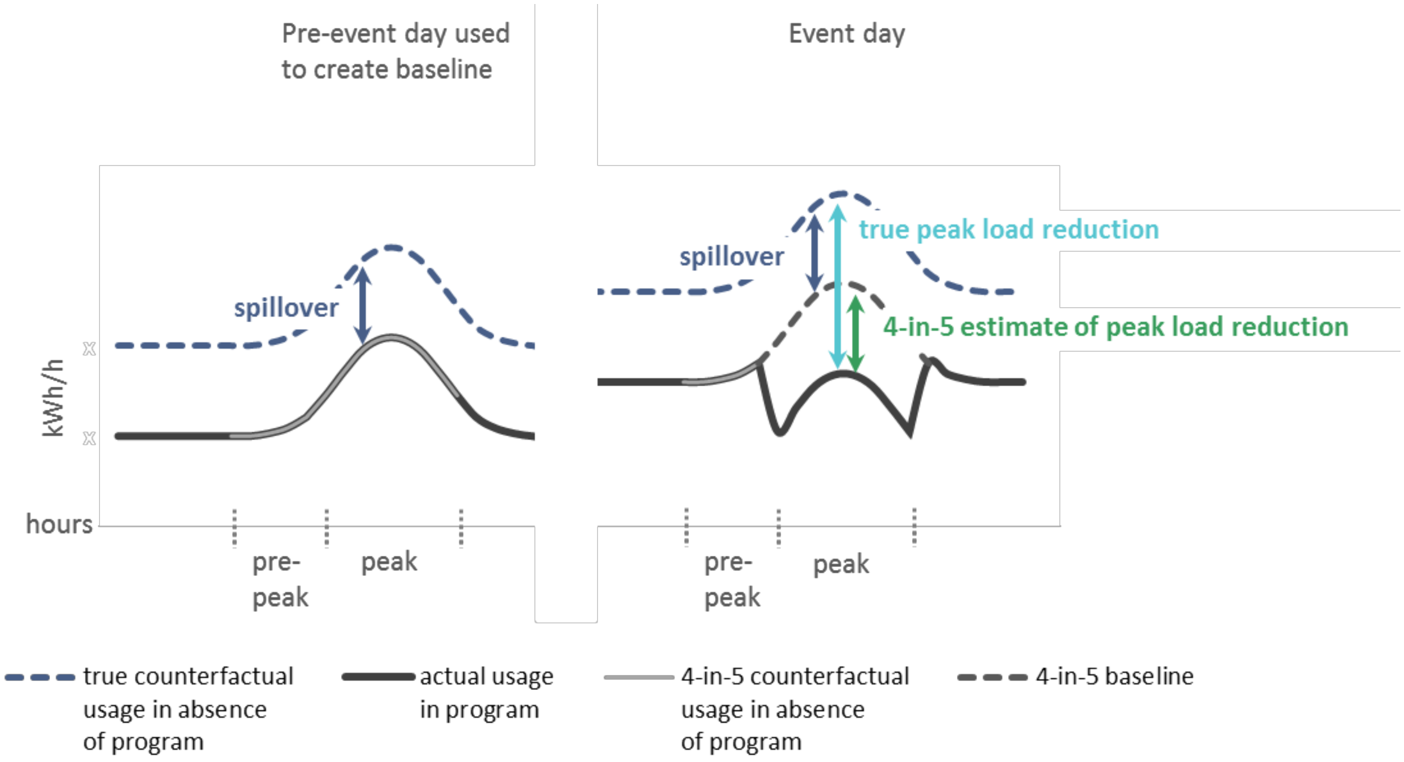


Figure 1. How Spillover affects the 4-in-5 Baseline and Peak Load Reduction Estimates

2.4. Theoretical Background on the Nature of Spillover

Our hypothesis that spillover is causing some of the total bias in the 4-in-5 method after adjusting for weather is rooted in theories from the Behavioral Sciences. Traditional economic theory predicts that customers would provide peak load reduction only during the higher priced hours (i.e., during peak hours on event days), and would increase usage during the lower priced hours (i.e., load shifting to off-peak hours and to non-event days). However, several theories from Behavioral Economics and Psychology, described in more detail below, suggest why this may be incorrect.

Limited attention: people have many small and/or large decisions to make every day. These decisions require mental energy and attention to think through and implement. People may choose to limit their attention to energy use decisions, so as not to use up precious mental energy (see, for example [13-15]). In this case, a change in energy behavior that only requires a one-time decision may be the best use of this limited attention. In the application of event-driven time-based rates or incentive-based programs, a customer may decide to respond by making a one-time decision to re-program their air conditioner thermostat to a higher temperature during peak hours of every day (or every week day), so that they don't have to think about making any incremental changes on the day an event is called. This means that the peak load would be reduced during peak hours on every day, not only during event days (i.e., spillover).

Habit formation: people are typically creatures of habit, with some set routines that do not change day-to-day (see, for example [16-18]). A household's evening routine may include a series of tasks that occur in the same order every day. For example: get home from work, make dinner, start the dishwasher, put the kids to bed, watch TV, go to sleep. Although this may be automatic and hard to change, a household might decide to make a conscious effort to form a new habitual routine. In our case, in response to participating in an event-driven time-based rate or incentive-based program, a customer may make an effort to change their habits in order to avoid more electric-intensive activities during peak hours. For example, the customer could change their routine to start the dishwasher only after peak hours (e.g., after watching TV rather than immediately after dinner). This would mean that peak load is reduced during peak hours on every day, not only on event days (i.e., spillover).

Risk aversion: people are typically risk averse (see, for example [19-23]). They do not like situations in which there is any risk of something bad happening, even if the possible gain is larger than the possible loss (for example, most people would not take a 50/50 bet of either winning $102 or losing $100, even though the possible gain of $102 is larger than the possible loss of $100). In the context of event-driven time-based rates or incentive-based programs, if a customer is worried that there is some risk they will not receive the event day notification (or will receive it but won't be home in time), it may seem safer to increase the programmed AC temperature on all days and incur slight discomfort even on non-event days, rather than risking a higher bill because they failed to respond to an event. These actions would again reduce the energy and demand that may have been used during non-event peak hours absent the program (i.e., spillover).

2.5. Quantifying the Bias from Spillover

To examine our hypothesis, we estimate spillover to determine if it exists and then quantify how it contributes to total bias in the 4-in-5 peak load reduction estimates for our CPP program. As shown in Eq. 3, we estimate spillover by determining the impact of the CPP rate using the control group from the RCT during the three baseline calculation periods (discussed in Section 2.1): (1) pre-peak hours on the event day (*SpilloverEventPrePeak*); (2) peak hours on baseline days (*SpilloverBaselinePeak*); and (3) pre-peak hours on baseline days (*SpilloverBaselinePrepeak*).

In order to quantify how spillover contributes to total bias, we estimate the "bias from spillover." We do this by calculating how spillover during the three baseline calculation periods affects the 4-in-5 baseline, and therefore biases the peak load reduction estimates. First we create a spillover-adjusted baseline that eliminates the effect of spillover for each of these periods. For each customer *i* and event *e*, where the *Spillover* variables are the estimated impacts of the CPP rate on usage during the three baseline calculation periods:

$$Spillover\ Adjusted\ Baseline_{ie} = \qquad \text{(Eq. 3)}$$

$$(BaselinePeak_{ie} - SpilloverBaselinePeak_{ie})$$

$$+(EventPrePeak_{ie} - SpilloverEventPrePeak_{ie})$$

$$-(BaselinePrePeak_{ie} - SpilloverBaselinePrePeak_{ie})$$

Comparing this spillover adjusted baseline to the baseline in Eq. 1 results in the bias from spillover, or the total amount that spillover affects the 4-in-5 estimates of peak load reduction:

$$\begin{aligned} Bias\ from\ Spillover_{ie} \\ &= SpilloverBaselinePeak_{ie} + SpilloverEventPrePeak_{ie} \\ &- SpilloverBaselinePrePeak_{ie} \end{aligned} \qquad \text{(Eq. 4)}$$

2.6. Alternative Baseline Methods

If spillover causes some of the total bias in the 4-in-5 method, a method that uses baseline calculation periods from *before* the program implementation (rather than during) may not be affected by spillover, and thus may be less biased.[7] Of course, these methods may suffer from other shortcomings; one example is that as customers remain on a time-based rate or incentive-based program for a longer amount of time, the pre-program baseline may become less relevant, and so a method that uses this technique may become less accurate over time.

We identified one such technique that derived a baseline using only pre-program time periods (hereafter referred to as LTAP method) [24]. The LTAP method is a simple linear regression baseline method. For each event, a baseline is created by taking a pre-program day with similar temperature, and then modified based on the difference in temperature between the event day and the pre-program day, as well as the relationship between temperature and usage (this relationship is estimated using a linear regression; more details are provided in Appendix B). We use the LTAP method to produce peak load reduction estimates, and then calculate its total bias using the method described in Section 2.2. Note that we only examine one alternative baseline method that is not affected by spillover; there may be many others.

2.7. Comparing Different Methods

In order to compare our two baseline methods, we calculate the root mean squared error (RMSE) of the estimates from the 4-in-5 method and the LTAP method relative to the estimates from the RCT method in order to have a quantitative measure of the overall accuracy (i.e., bias and precision) of these methods across all events. We compare the RMSE metric of the LTAP method to that of the 4-in-5 method. Within each method, we also compare the RMSE of estimates for the default enrollment approach to that of the voluntary enrollment approach.

3. SMUD's Consumer Behavior Study

The randomized design of SMUD's consumer behavior study provides a unique opportunity to examine the total bias and proportion of the bias resulting from spillover of the 4-in-5 baseline method. SMUD conducted one of the largest and most extensive consumer behavior studies under the SGIG program [25].[8] Like most of the other consumer behavior studies implemented under the SGIG program, SMUD's study utilized a true experimental design (i.e., randomized control trial and randomized encouragement design) in order to more credibly and precisely estimate the load response to these various rates [8].

One of the study's main goals was to better understand how the enrollment approach (voluntary vs. default) affected participation rates, drop-out rates, and electricity demand impacts in response to different time-based rate designs in effect during the summer months (June through September) of 2012 and 2013. These included:

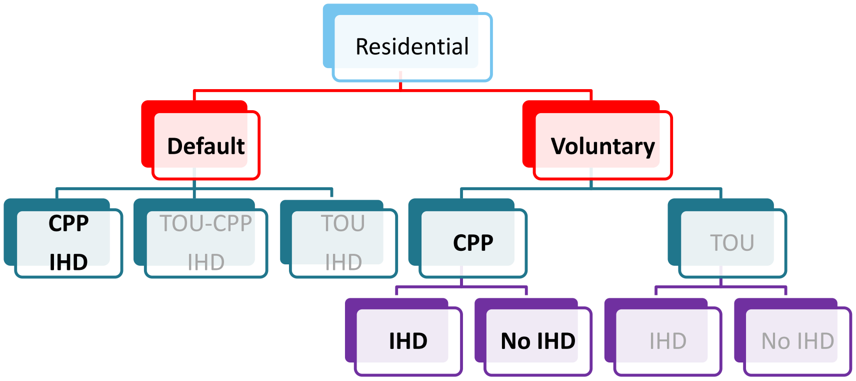
(1) A two-period TOU rate with a three-hour (4-7 p.m.) peak period;

[7] Although even in this case there may be spillover; customers may learn that the program is going to go into effect and modify their behavior before it begins.

[8] For more details on SMUD's consumer behavior study, see Appendix A.

(2) CPP overlaid on SMUD's standard inclining block rate; and
(3) CPP overlaid on the study's TOU rate.

The study's design is summarized in Figure 2 and the rates are presented in Table 1. The CPP rate was designed and implemented with 12 critical events called each year between the hours of 4 PM and 7 PM (i.e., 48 hours in total) on summer weekdays, excluding holidays. For the purposes of this analysis, only the customers included in the CPP overlaid on the underlying inclining block rate, including both enrollment approaches (i.e., voluntary, default) and technology treatments (i.e., with or without the presence of an in-home display (IHD) offer) were analyzed and discussed. Specifically, the four randomized groups we examine are: (1) voluntary CPP with IHD, (2) voluntary CPP with no IHD, (3) default CPP with IHD, and (4) a control group on SMUD's standard inclining block rate.



Note: Those treatment arms depicted in gray were not analyzed here, while those with black text were included in this study.

Figure 2. SMUD's Consumer Behavior Study Experimental Design

Table 1. SMUD's CBS Summer 2012 Rate Design (¢/kWh)[9]

Period	CPP in ¢/kWh (Treatment)	Inclining Block in ¢/kWh (Control)
Non Critical Peak Base (< 700 kWh)	8.51	9.38
Non Critical Peak Base-Plus (> 700 kWh)	16.65	17.65
Critical Peak	75.0	N/A

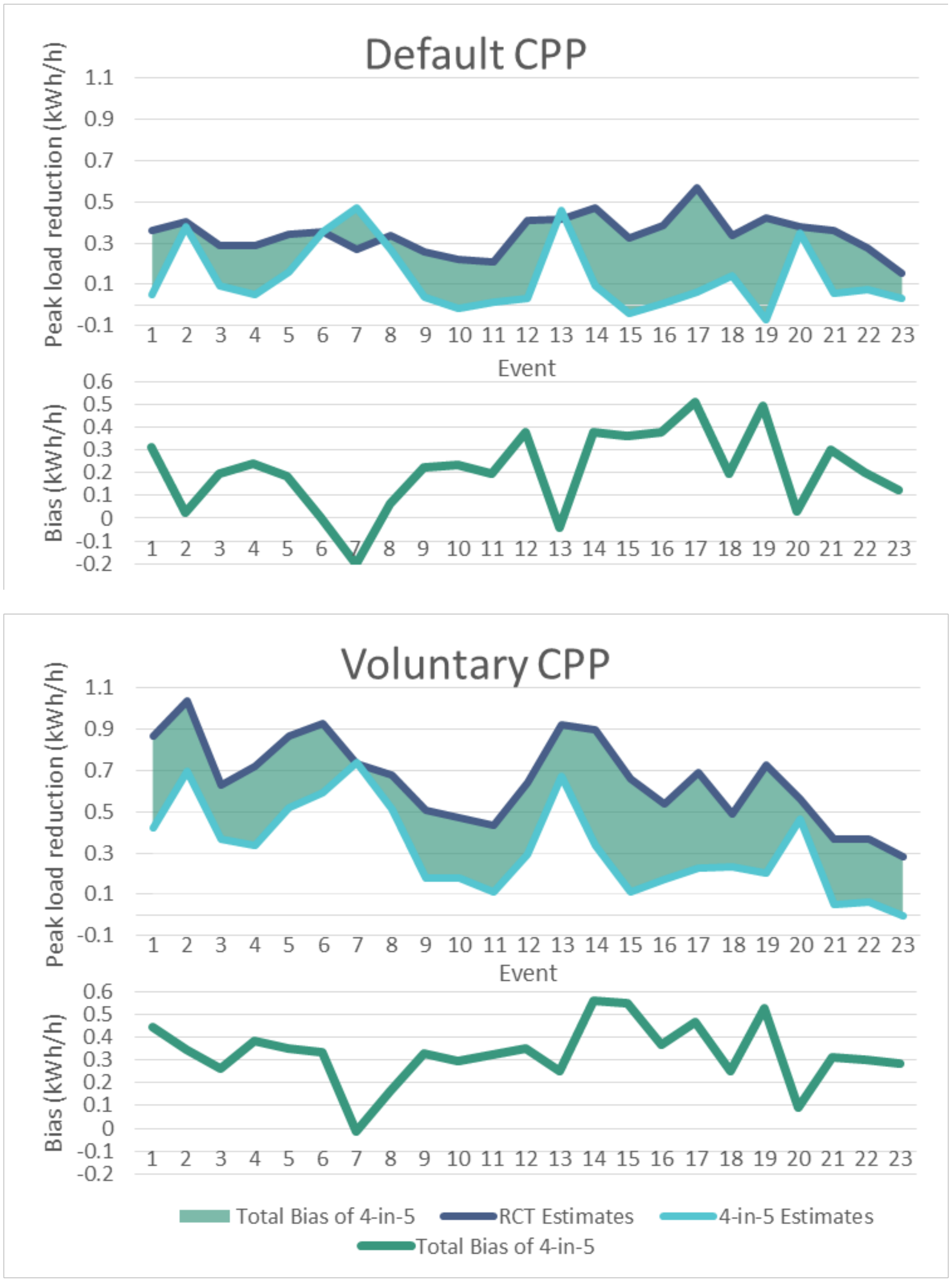
[9] Table 1 shows the rates charged to SMUD's general population of customers on the CPP treatment rates. SMUD also included customers enrolled in the low-income rate, referred to as Energy Assistance Program (EAPR). These customers faced a lower fixed charge than non-EAPR customers, and were given a discount of 35% applied to electricity use charges for base use, and a discount of 30% applied to non-base use up to 600kWh, above which no discount was applied. This same discount structure applied to both time-based treatment rates and inclining block standard rates.

4. Results

4.1. Total Bias Calculations for the 4-in-5 Method

As described previously, the total bias associated with the 4-in-5 method is the difference between the peak load reduction estimates generated using the RCT and those generated using the 4-in-5 baseline. Our results show that for the CPP program we are examining, the 4-in-5 method is systematically biased (i.e., consistently over- or under-estimates the reduction), as seen in Figure 3. For both voluntary and default enrollment approaches, total bias appears to exist in almost every event, almost always underestimates the peak load reduction, and is of a meaningful magnitude. Most of these estimates of bias are statistically significant (see Appendix B for the full regression results). For the default enrollment approach, the total estimated bias ranges between -0.20kWh and 0.51kWh per hour across different events, with an average estimated bias of 0.21kWh per hour across all events. For the voluntary enrollment approach, the total estimated bias ranges between -0.10kWh and 0.56kWh across different events, with an average estimated bias of 0.33kWh across all events. The 4-in-5 method overestimates the peak load reduction in only two out of 23 events for the default enrollment, and only one event for the voluntary enrollment, with the vast majority of cases resulting in an underestimate. The underestimate is substantial: on average, the 4-in-5 method produces savings estimates that are 39% and 46% of the unbiased RCT estimates for the default and voluntary enrollment approaches, respectively. The RMSE overall accuracy metric is 0.27 for the default enrollment approach, and 0.35 for the voluntary enrollment approach; by this metric, the 4-in-5 method is slightly more accurate for the default approach.

One interesting finding is that for both the default enrollment and voluntary enrollment, the bias appears to be lower for a handful of specific events (i.e., 7, 13, and 20). There are many possibilities for why this may be the case. For example, it may be that those events had variable weather patterns that were dissimilar to the hours and/or days ahead of the event used to calculate the baseline, so that weather was not fully controlled for and therefore the estimates were more biased on those days. It may also be that the baseline method is more biased during certain days of the week and less on others, or that the baseline may not be as accurate for an event that is closer to another event.. To unpack this, we would need more observations of events so that we could perform a regression that tests the effects of these potential factors (as well as others) on the bias of each event.



Notes: Y axis shows kWh/h on average across all participants for each event.

Figure 3. Total Bias in Peak Load Reduction Estimates for the 4-in-5 Baseline Method

4.2. Quantification of the Bias from Spillover

Our results in Section 4.1 are consistent with our theory that spillover is causing systematic bias: the peak load estimates from the 4-in-5 baseline method tend to underestimate the reductions relative to the RCT estimates. We now directly test this hypothesis by estimating spillover and quantifying the proportion of the estimated total bias that is due to spillover.

Our results are shown in Figure 4. For each graph, the top portion shows the estimated bias due to spillover. For reference, the top portion also includes the results from Section 4.1 on the total bias of the 4-in-5 method. The bottom portion depicts the percent of the total bias that can be explained by the spillover.[10]

We find evidence of spillover onto non-event days: the spillover estimates are statistically significant during peak hours on the pre-event days used as a baseline calculation period.[11] However, during pre-peak hours on event days and on pre-event days, spillover is not statistically significant but is still non-zero.[12] Despite the lack of statistical significance for bias estimates in two of the three time periods, these estimates still represent our best guess for the size of the actual spillover. We therefore use all of three estimates when we combine the spillover during each of these three periods in order to calculate the proportion of total bias due to spillover.

There are a few interesting observations to draw from our results.

First, as shown in the top part of the figures, bias from spillover appears to exist during every event. For the default enrollment approach, the estimated bias from spillover ranges between 0.04kWh and 0.17kWh per hour across events, with an average bias from spillover of 0.12kWh per hour. For the voluntary enrollment approach, the estimated bias from spillover ranges between 0.15kWh and 0.36kWh per hour across events, with an average bias from spillover of 0.26kWh per hour.

Second, although spillover contributes to the total bias in 4-in-5 estimates, it clearly is not the only factor causing bias; the bias from spillover cannot explain the full amount of the total bias in the 4-in-5 estimates. For the default enrollment approach, the portion of total bias explained by spillover ranges between 21% and 83% across all events, with an average of 47%. For the voluntary enrollment approach, the portion of total bias explained by spillover ranges between 36% and 97% across all events, with an average of 65%. In addition, there are events for which the total bias goes in the *opposite* direction of the bias from spillover; the bias from spillover is always positive, but the total bias is negative for a few events. This means that for those events, although spillover is pulling the total bias in one direction, there are additional factors that are pulling the total bias in the other direction. Future research could examine and identify other sources of specific bias to determine if together with spillover they fully explain the total bias.

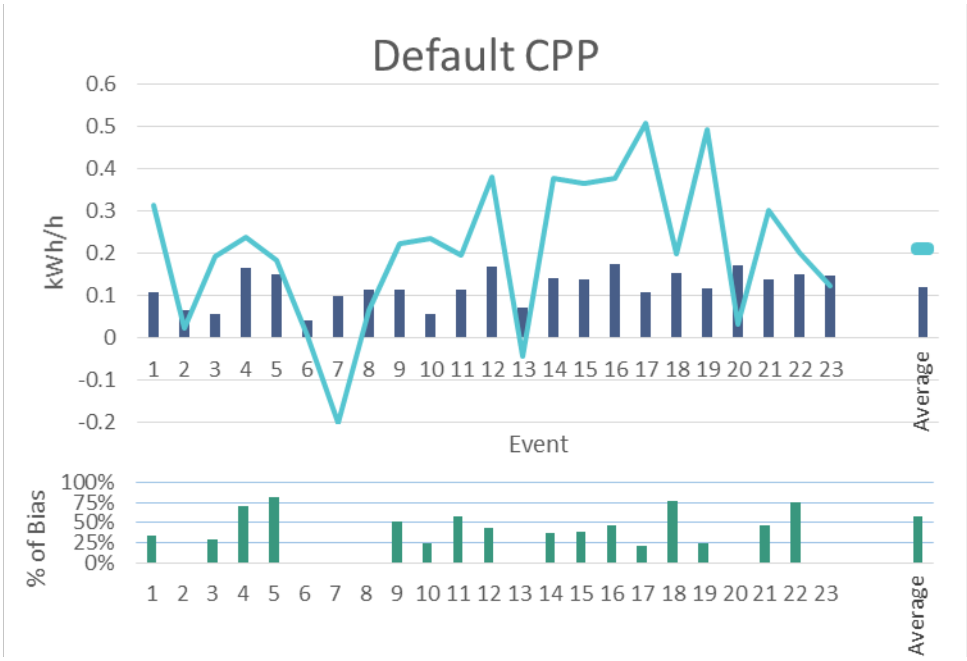
Third, the estimated bias from spillover is greater for the voluntary versus the default enrollment approach: 0.26kWh versus 0.12kWh. This is consistent with the conjecture made in Section 2.4 that customers who take action to opt-in to a voluntary rate may be more engaged, interested, and knowledgeable about their own energy choices than those who are placed on the rate by default. These voluntary participants may therefore be more motivated and willing to change their energy behavior both during the event and during non-event hours. This would lead to greater bias from spillover in savings estimate using the 4-in-5 method for the voluntary enrollment approach.

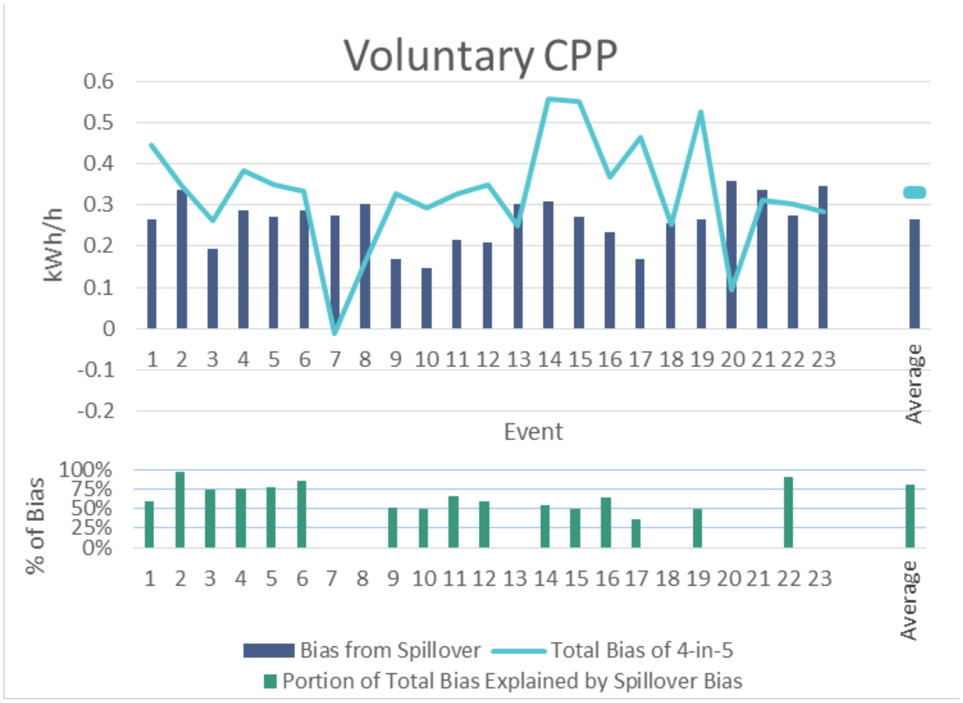
[10] Negative percentages are excluded.

[11] See Appendix B for details on estimates.

[12] This may be because the effect is smaller during off-peak hours, and our sample size is not large enough to detect an effect.

The bias from spillover also accounts for a greater *portion* of the total bias in the voluntary versus the default enrollment approach: 81% of the total bias is explained by the spillover for voluntary versus 58% for default.[13] This means that a greater portion of the total bias is left unaccounted for in the default enrollment approach.





[13] This is not an obvious finding; although it might be expected that the peak load reduction, spillover, and bias is larger for the voluntary rate than the default rate, one might expect these to scale proportionally, so that while more spillover causes more bias, the spillover contributes the same *percentage* to the bias. However, such is not the case.

Note: Y axis of top graph shows kWh/h on average across all participants for each event.

Figure 4. Bias from Spillover and Portion of Total Bias Explained by Spillover

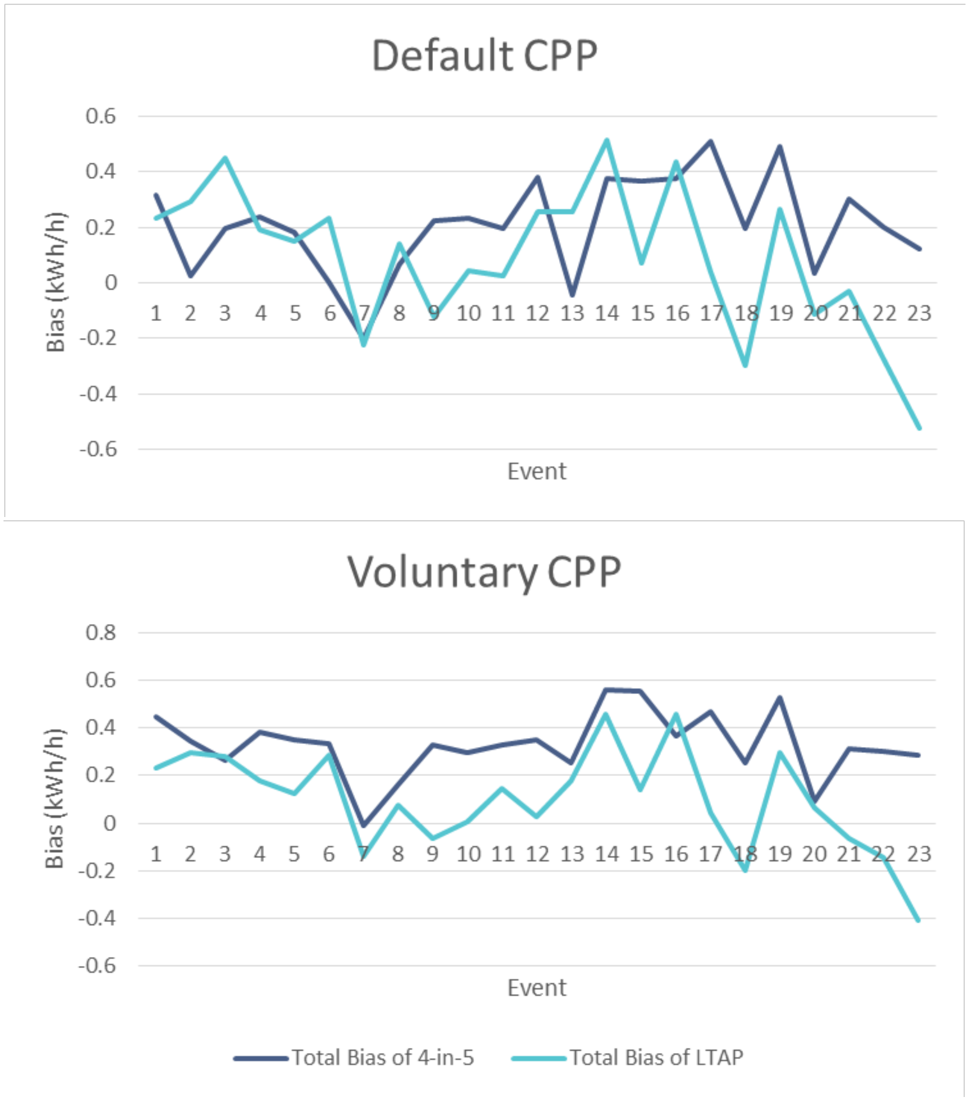
4.3. Total Bias Calculations of Alternative Baseline Methods

We have shown that spillover causes part of the total bias in the 4-in-5 method. We now turn to estimating load impacts from a method that is hypothesized to have less susceptibility to bias coming from spillover: the LTAP method (this method calculates a baseline using only pre-program time periods and is therefore not as affected by spillover, as discussed in Section 2.6).

Figure 5 shows the bias estimates from the LTAP method compared to bias estimates from the 4-in-5 method. First, we use the RMSE metric to examine the overall accuracy of each method. Using this metric, the LTAP method appears to be more accurate for the voluntary enrollment approach, but not the default approach. For the default enrollment approach, the RMSE is 0.27 for both the 4-in-5 and LTAP; by this accuracy metric, both methods perform the same. For the voluntary enrollment approach, the RMSE is 0.35 for 4-in-5 and 0.28 for LTAP; by this metric LTAP does better (i.e., is more accurate overall).[14]

Next we examine the systematic bias (i.e., whether the estimate consistently over- or under-estimates the peak load reductions). The LTAP appears to be systematically biased, but less so than the 4-in-5 method. For the default approach, the 4-in-5 method underestimates the peak load reductions in 91% of the events, while the LTAP underestimates only 70% of the time; for the voluntary approach, the 4-in-5 method underestimates the peak load reductions in 96% of the events, while the LTAP underestimates only 74% of the time. Because the LTAP is less systematically biased, it has a lower estimated total bias when averaged across all events: for the default enrollment approach, the average estimated total bias is 0.09kWh for the LTAP versus 0.21kWh for 4-in-5; for the voluntary enrollment approach, the average estimated total bias is 0.10kWh for the LTAP versus 0.33kWh for 4-in-5.

[14] Another observation is that the LTAP method appears to perform worse in the last few events. Although we cannot draw any conclusions from so few data points, one possible explanation is that because the LTAP method is based on usage from before the rate begins, it becomes less accurate over time (as previously discussed).



Notes: Y axis shows kWh/h on average across all participants for each event. Differences between the LTAP and the 4-in-5 method are not statistically significant.

Figure 5. Total Bias in Peak Load Reduction Estimates for the LTAP and 4-in-5 Methods

5. Discussion & Conclusion

As the penetration of time-based rates and incentive-based programs expand, the need to produce accurate and unbiased estimates of the peak load reductions resulting from these programs will be increasingly important. Previous research has shown that estimates generated from 4-in-5 baseline methods, identified as the best performing baseline method [4, 5], are biased [10]. In this paper, we quantify this bias for a CPP program leveraging its RCT experimental design and look deeper into the cause and implications of this bias.

Our results show that there is spillover of electricity reduction from the higher priced hours targeted by the program onto other time periods. We show that this spillover can explain some, but not all, of the bias in the 4-in-5 method. Because the hours not targeted by the program have slightly lower rates, this means that customers actually *reduced* their usage during lower priced hours, contrary to what traditional economic theory would predict. This suggests that elements from Behavioral Economics and Psychology, such as limited attention, habit formation, and risk aversion, may help explain a customer's decision making process. Further research could examine whether and how much each of these, or other behavioral factors, are responsible for spillover.

Research is currently underway to develop new baseline methods that are less biased and more accurate. We examined one such method, a novel baseline method called LTAP. Although this method is not affected by bias due to spillover, and did perform just as well or better than the 4-in-5 method with respect to a metric of overall accuracy (i.e., RMSE), its estimates were still somewhat inaccurate.

Researchers should continue to develop new evaluation methods which are less biased than existing methods, given the myriad of challenges the electric industry has faced in implementing gold standard RCT experimental designs. One potential improvement is using more complex analytical techniques in baseline methods, which have historically been very simplistic. Recent research in other fields has demonstrated the value of using sophisticated machine learning algorithms and advanced econometrics techniques. These include: boosting, classification, and regression tree algorithms, such as random forests, which can be used to estimate a more accurate propensity score for an econometrics matching method; incorporating algorithms such as LASSO, boosting, and regression trees can improve the accuracy of estimates; using a machine learning bayesian structural time series can produce a more accurate prediction of the energy usage of a household in the absence of the treatment for use as a control group, and finally an advanced econometric method called regression discontinuity has been found to perform better than traditional matching methods (for discussions on using these methods, see [12, 26-32]).[15] Future research could quantify the reduction in bias due to the application of these cutting edge analytical techniques in demand response program evaluations.

From a policy standpoint, our findings also have broader implications for demand side management portfolio planning. The reduction in electricity consumption during hours not targeted by the DR opportunity (i.e., spillover) suggests that this form of event-driven DR could contribute to overall energy savings goals (e.g., energy efficiency portfolio standards). If utilities take a more holistic approach to DSM portfolio planning, such goals could likely be realized at lower cost.[16]

Understanding the cause of the spillover also has important implications for the optimal design and implementation of DR programs. For example, if spillover is caused by customers reacting to an event-driven DR program by re-programming their thermostats during all weekdays, then there will be reductions in peak demand during non-event days as well as event days. However, if an event-driven DR

[15] One should not blindly pursue greater complexity, presuming it always results in greater accuracy. For example, overfitting an econometric model so that it appears to be accurate given past data but actually fails at predicting the future is just as big of a problem as applying a methodology that both inaccurately explains past data and poorly predicts the future. Recently, data scientists have begun to develop machine learning techniques to deal with this issue (e.g., dividing the data into separate sets for the purpose of training, testing, and validation [28]). Being aware of the both the challenges introduced by greater complexity as well as the potential approaches for mitigating those challenges should help our industry move forward with better and more accurate baselines.

[16] Some utilities, including PG&E [33] are currently working on integrating the DR and EE benefits of Behavioral Demand Response Programs.

program uses automated controls that respond exclusively to events, this automation may improve response during events but decrease peak reductions on non-event days.

In addition, DR opportunities may currently be undervalued when assessing their cost effectiveness. Typically the only benefit taken into account for DR programs like these are the reduction in demand during specific targeted hours, rather than including the benefit of overall energy savings outside of those targeted hours [8]. Future research should explore potential synergies of DR opportunities with other DSM programs to ascertain if policies should be modified to promote a more integrated approach to demand side management.

Acknowledgements

The work described in this study was funded by the U.S. Department of Energy's Office of Electricity Delivery and Energy Reliability under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

The authors would like to thank Joe Paladino (DOE OE) for his ongoing support. In addition, the authors would like to thank the following individuals for reviewing a draft of this manuscript: Ken Agnew (DNV GL), Patrick Baylis (Stanford University), Miriam Goldberg (DNV GL), Doris Lee (Student), Olivia Patterson (Opinion Dynamics), Steve Schiller (LBNL), Alexander Sim (LBNL), Elizabeth Stuart (LBNL), and K. John Wu (LBNL).

Appendix A: Background on SMUD's SGIG Consumer Behavior Study

Sacramento Municipal Utility District (SMUD) is a summer peaking municipal electric utility with ~625,000 customers in its ~900 square mile service territory that covers much of the Sacramento, CA metropolitan area. SMUD's SGIG project (SmartSacramento) includes a consumer behavior study that evaluates customer acceptance and response to enabling technology combined with various time-based rates under different recruitment methods. The utility is targeting AMI-enabled residential customers across the entire service territory to participate in the study.

This study focuses on evaluating the timing and magnitude of changes in residential customers' peak demand patterns due to exposure to varying combinations of enabling technology, different recruitment methods (i.e., opt-in vs. opt-out), and several time-based rates. SMUD is also interested in learning about customer acceptance of the different time-based rates under the alternative recruitment methods.

Rate treatments include the implementation of three time-based rate programs in effect from June through September: a two-period TOU rate that includes a three-hour on-peak period (4 - 7 p.m.) each non-holiday weekday; a CPP overlaid on their underlying tiered rate; and a TOU with CPP overlay (TOU w/CPP) (see Table A-1). Customers participating in any CPP rate treatments receive day-ahead notice of critical peak events, called when wholesale market prices are expected to be very high and/or when system emergency conditions are anticipated to arise. CPP participants will be exposed to 12 critical peak events during each year of the study.

Control/information technology treatments include the deployment of IHDs. SMUD is offering IHDs to all opt-out customers in any given treatment group and to more than half of the opt-in customers in the treatment group. All participating customers receive web portal access, customer support and a variety of education materials.

Table A-1. SMUD CBS Rate Design (¢/kWh)

Period	CPP	TOU	TOU-CPP
Base (< 700 kWh)	8.51		
Base (> 700 kWh)	16.65		
Off-Peak (< 700 kWh)		8.46	7.21
Off-Peak (>700 kWh)		16.60	14.11
Peak		27.00	27.00
Critical Peak	75.00		75.00

Due to the variety of treatments, the study includes three different experimental designs: randomized controlled trial (RCT) with delayed treatment for the control group, randomized encouragement design (RED) and within-subjects design (see Figure A-1).

In all three cases, AMI-enabled residential customers in SMUD's service territory are initially screened for eligibility and then randomly assigned to one of the seven treatments or the RED control group.

For the two treatments that are included in the RCT "Recruit and Delay" study design, customers receive an invitation to opt in to the study where participating customers receive an offer for a specific treatment. Upon agreeing to join the study, customers are told if they are to begin receiving the rate in the first year of the study (i.e., June 2012) or in the summer after the study is complete (i.e., June 2014).

For two of the three treatments that are included in the RED, customers are told that they have been assigned to a specific identified treatment but have the ability to opt out of this offer. Those who do not opt out receive the indicated treatment for the duration of the study. Those who opt out are nonetheless included in the study's evaluation effort but do not receive the indicated treatment. For one of the three RED treatments, customers receive an invitation to opt in to the study where participating customers receive a specific treatment. Customers that opt in are then assigned to receive the treatment in year 1 of the study (i.e., 2012).

For the two treatments that are included in the within-subject design, customers are told they have been assigned to either the Block w/CPP treatment or the TOU w/CPP treatment with technology.[17] In the former case, customers only have the ability to opt in to this specific treatment. In the latter case, customers only have the ability to opt out of this specific treatment.

[17] The within-subjects method was designed to use no explicit control group; instead it estimates the effects of the treatment for each participant individually, using observed electricity consumption behavior both before and after becoming a participant in the study as well as on critical peak event and non-event days. However, the control group selected for the RED design may be used as a control group.

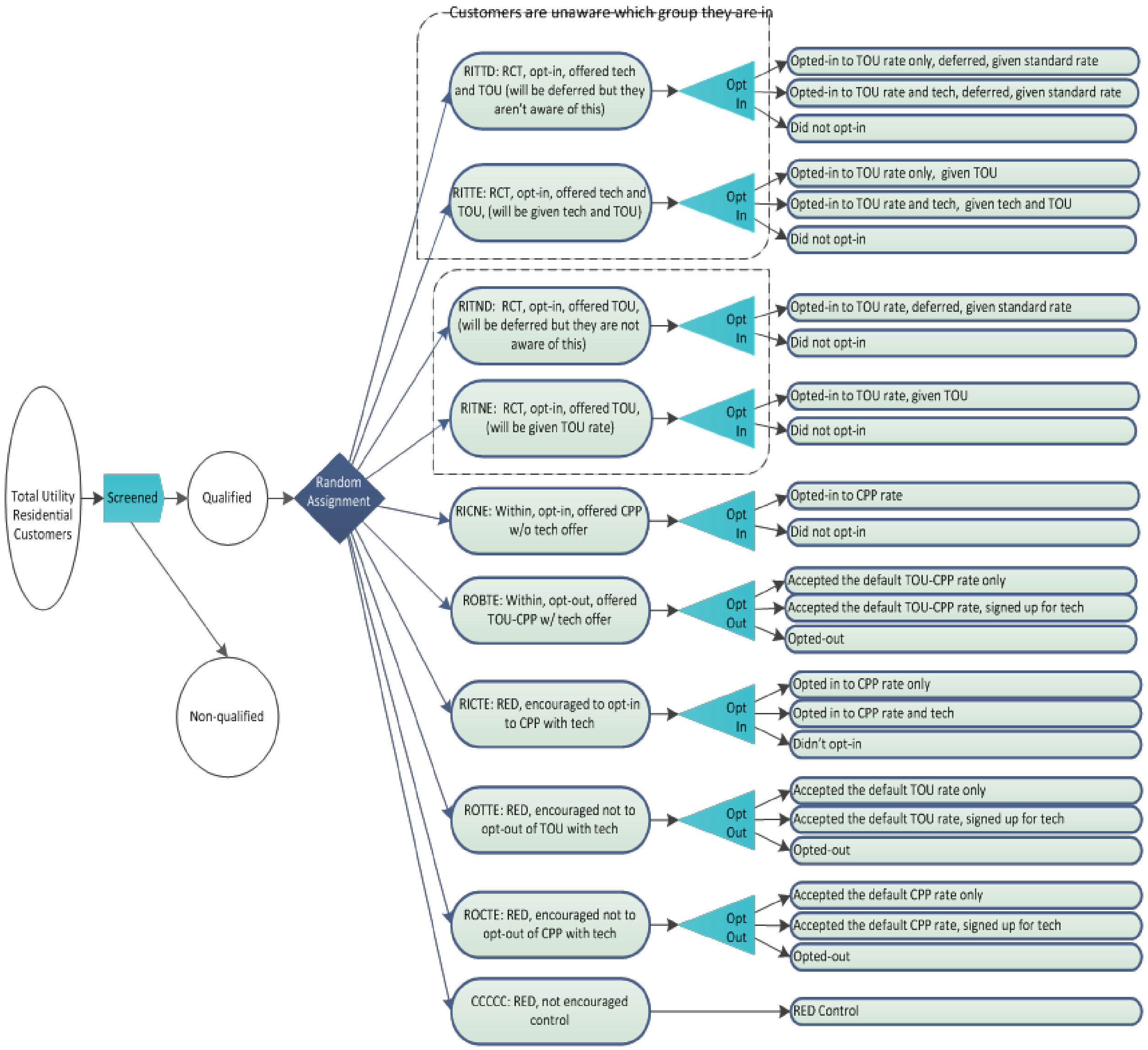

Customers are unaware which group they are in
Total Utility Residential Customers
Screened
Qualified
Non-qualified
Random Assignment
RITTD: RCT, opt-in, offered tech and TOU (will be deferred but they aren't aware of this)
RITTE: RCT, opt-in, offered tech and TOU, (will be given tech and TOU)
RITND: RCT, opt-in, offered TOU, (will be deferred but they are not aware of this)
RITNE: RCT, opt-in, offered TOU, (will be given TOU rate)
RICNE: Within, opt-in, offered CPP w/o tech offer
ROBTE: Within, opt-out, offered TOU-CPP w/ tech offer
RICTE: RED, encouraged to opt-in to CPP with tech
ROTTE: RED, encouraged not to opt-out of TOU with tech
ROCTE: RED, encouraged not to opt-out of CPP with tech
CCCCC: RED, not encouraged control
Opt In
Opt Out
Opted-in to TOU rate only, deferred, given standard rate
Opted-in to TOU rate and tech, deferred, given standard rate
Did not opt-in
Opted-in to TOU rate only, given TOU
Opted-in to TOU rate and tech, given tech and TOU
Did not opt-in
Opted-in to TOU rate, deferred, given standard rate
Did not opt-in
Opted-in to TOU rate, given TOU
Did not opt-in
Opted-in to CPP rate
Did not opt-in
Accepted the default TOU-CPP rate only
Accepted the default TOU-CPP rate, signed up for tech
Opted-out
Opted in to CPP rate only
Opted in to CPP rate and tech
Didn't opt-in
Accepted the default TOU rate only
Accepted the default TOU rate, signed up for tech
Opted-out
Accepted the default CPP rate only
Accepted the default CPP rate, signed up for tech
Opted-out
RED Control

Figure A-1. SMUD Recruitment Process

Appendix B: Data Analysis Methods and Results

Peak Load Reduction Estimates and Bias Calculations

RCT Method

Table B-1 shows the results of the peak load reduction estimates and the calculated bias for the RCT method. The average peak period load impacts estimates for the two treatment groups (Default and Voluntary) were estimated using a difference-in-differences (DID) instrumental variables (IV) regression using Two-Stage Least Squares (2SLS). While whether or not a household actually experiences the study TOU electricity rates is not random (because of self-selection in or out of treatment), the assignment to a treatment group is random. We can therefore use *assignment* to treatment (or "encouragement" as it's known in the literature) as an instrument for *actual* treatment (i.e., exposure to the treatment time-of-use rate).

A separate regression is run for each treatment group (Default or Voluntary). We instrument for T_{it} with randomized assignment (or encouragement) to treatment indicator A_{it}.

$$T_{it} = \delta A_{it} + \gamma_i + \tau_t + e_{it} \qquad \text{(Eq. B-1)}$$

T_{it} is an indicator variable is equal to one starting on June 1st, 2012 if household *i* was actually enrolled in treatment and remained in the treatment group at time *t*, zero otherwise. A_{it} is an indicator variable equal to one starting on June 1st, 2012 if household I was encouraged to be in one of the treatment groups (random assignment to treatment), zero otherwise. The predicted values $\hat{T}_{it}$ are then used in Eq. B-2.

The estimating equation we use to derive the estimates in Table B-1 is as follows:

$$y_{it} = \beta\hat{T}_{it} + \gamma_i + \tau_t + \varepsilon_{it} \qquad \text{(Eq. B-2)}$$

The variable y_{it} is hourly electricity consumption for household *i* in hour *t*; $\hat{T}_{it}$ are the predicted values generated from the regression shown in equation (1); γ_i is a household fixed effect;[18] τ_t is an hour of sample fixed effect[19]; and ε_{it} is the error term assumed to be distributed IID normal across households. In order to account for serial correlation across time observations within households, we clustered the standard errors of the estimates at the household level. The data used are peak hour consumption (4 pm to 7 pm) on non-holiday weekdays in both treatment summers (2012 and 2013) and in the pre-treatment summer (2011). Households in both the treatment groups and the control group are included. Coefficient β captures the average hourly treatment effect per household.

4-in-5 Method

Table B-1 also shows the results of the peak load reduction estimates and the calculated bias for the 4-in-5 method. The estimates are calculated as described in Section 2.2, and the total bias associated with the 4-in-5 method is the difference between the peak load reduction estimates generated using the RCT and those generated using the 4-in-5 baseline.

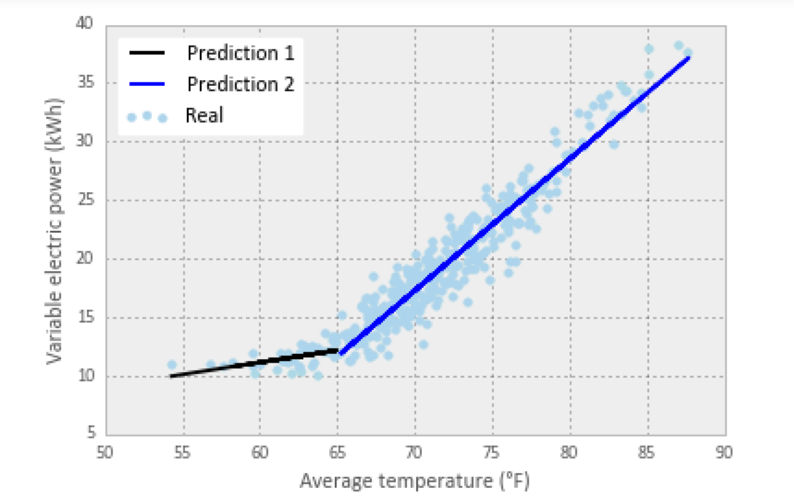
Linear Relation Between Temperature and Aggregated Power (LTAP)

Table B-1 also shows the results of the peak load reduction estimates and the calculated bias for the LTAP method. LTAP, as described in Kim et al [24], is a white-box model which assumes a linear relationship between the aggregate household electricity usage and the outdoor temperature, as shown in Figure B-1. Intuitively, when the outdoor temperature increases, the indoor temperature also increases after a delay due to insulation. When the temperature exceeds a threshold, the households

[18] In the tables that follow which show the output from the econometric analysis, the row titled "Household Fixed Effects" with a value of "Yes" indicates when these household-level fixed effects were applied.

[19] In the tables that follow which show the output from the econometric analysis, the row titled "Hour of Sample Fixed Effects" with a value of "Yes" indicates when these hour of sample fixed effects were applied.

turn on their air-conditioners. The *variable electricity usage* (v_0) is defined as the usage minus the *base load* (b_0) from electrical appliances that require constant electricity usage (e.g., refrigerators and water pumps). The higher indoor temperature causes the air-conditioners to consume more energy to remove the heat. Therefore, the variable electricity usage has a strong linear relationship with the outdoor temperature, while the base load is constant.



Note: Linear dependence between outdoor average temperature and variable electricity usage (R=0.95). Source: Kim et al. [24].

Figure B-1. Linear Regression of Two Piecewise Linear Spline Function.

For the analysis presented in Section 2.6, LTAP first performs a linear regression over the aggregate customer usage data. The computed regression coefficients are then used for predicting the aggregate daily usage (a_1) for each customer based on their total daily usage on the *reference day* (the day with the closest temperature in year T-1) (a_0).

$$a_1 = a_0 + s(t_1 - t_0).$$

(Eq. B-3)

To predict the hourly usages (h_1), we assume the household's daily profile remains the same during the study period. We scale the variable load hourly usage based on the predicted and reference day total usage as follows:

$$h_1[i] = b_0 + (h_0[i] - b_0)a_1/a_0.$$

(Eq. B-4)

LTAP baseline model relies on two key assumptions: the aggregate daily usage depends linearly on outdoor temperature, and each household's usage profile stays the same across the three years of this study. Both of these assumptions could be violated when the household changes in some way, such as adding new electrical appliances, changing the number of occupants, adding new insulation and so on. The impact of these factors needs to be studied with more relevant data.

From our current study, we see that LTAP has many advantages over existing time-series predictions models, which often require past prediction of nearby dates. For example, when predicting usage for July 6th in year T+1, the model requires the usages of a few days before, such as those on July 5th and June 29th in year T+1. Since usage values from July 5th and June 29th in year T+1 have to be predicted first before used, there is a strong possibility the prediction errors would accumulate over time, which can make long-term predictions highly unreliable. LTAP mitigates this effect by using the historical usage on the reference day without any intermediate prediction steps. Therefore, LTAP has a greater potential for making long-term, stable prediction, which is useful for planning and building infrastructure to support the power demand in a local region.

The total bias associated with the LTAP method is the difference between the peak load reduction estimates generated using the RCT and those generated using the LTAP.

Table B-1. Peak Load Reduction Estimates for RCT, 4-in-5 Method, and LTAP Method, and Bias of the 4-in-5 Method and LTAP Method

Default CPP

Event	**RCT**	*std error*	**4-in-5**	*std error*	**LTAP**	*std error*	**Total Bias of 4-in-5**	**Total Bias of LTAP**
1	0.363	*0.056*	0.048	*0.052*	0.131	*0.072*	0.315	0.232
2	0.404	*0.061*	0.380	*0.055*	0.111	*0.073*	0.024	0.293
3	0.287	*0.057*	0.093	*0.047*	-0.164	*0.070*	0.194	0.451
4	0.287	*0.055*	0.049	*0.050*	0.096	*0.065*	0.238	0.190
5	0.345	*0.057*	0.161	*0.054*	0.193	*0.072*	0.183	0.152
6	0.357	*0.060*	0.353	*0.055*	0.124	*0.075*	0.004	0.233
7	0.272	*0.053*	0.473	*0.052*	0.493	*0.065*	-0.201	-0.222
8	0.334	*0.052*	0.268	*0.051*	0.192	*0.069*	0.066	0.143
9	0.261	*0.051*	0.037	*0.040*	0.384	*0.060*	0.223	-0.124
10	0.218	*0.053*	-0.016	*0.048*	0.176	*0.061*	0.235	0.042
11	0.207	*0.052*	0.013	*0.045*	0.181	*0.064*	0.194	0.026
12	0.413	*0.068*	0.031	*0.056*	0.155	*0.083*	0.382	0.258
13	0.418	*0.064*	0.462	*0.050*	0.164	*0.080*	-0.044	0.254
14	0.469	*0.072*	0.092	*0.058*	-0.043	*0.081*	0.377	0.512
15	0.325	*0.061*	-0.039	*0.054*	0.253	*0.069*	0.364	0.072
16	0.384	*0.064*	0.007	*0.045*	-0.052	*0.066*	0.376	0.436
17	0.572	*0.069*	0.063	*0.055*	0.531	*0.080*	0.509	0.041
18	0.337	*0.059*	0.139	*0.038*	0.633	*0.062*	0.198	-0.296
19	0.421	*0.064*	-0.071	*0.053*	0.156	*0.069*	0.492	0.265
20	0.382	*0.059*	0.350	*0.040*	0.494	*0.064*	0.033	-0.112
21	0.360	*0.062*	0.059	*0.040*	0.390	*0.060*	0.301	-0.030
22	0.277	*0.063*	0.076	*0.032*	0.553	*0.055*	0.201	-0.276
23	0.152	*0.064*	0.031	*0.027*	0.676	*0.056*	0.122	-0.523

Voluntary CPP

Event	**RCT**	*std error*	**4-in-5**	*std error*	**LTAP**	*std error*	**Total Bias of 4-in-5**	**Total Bias of LTAP**
1	0.867	*0.076*	0.422	*0.040*	0.638	*0.052*	0.445	0.229
2	1.040	*0.086*	0.694	*0.043*	0.743	*0.054*	0.346	0.296
3	0.629	*0.077*	0.367	*0.038*	0.351	*0.051*	0.261	0.277
4	0.719	*0.074*	0.336	*0.034*	0.543	*0.047*	0.383	0.176
5	0.871	*0.081*	0.520	*0.040*	0.745	*0.052*	0.351	0.125
6	0.930	*0.086*	0.596	*0.043*	0.645	*0.053*	0.334	0.285
7	0.731	*0.073*	0.742	*0.039*	0.869	*0.052*	-0.011	-0.138
8	0.677	*0.076*	0.513	*0.039*	0.602	*0.049*	0.163	0.075
9	0.507	*0.075*	0.179	*0.029*	0.569	*0.041*	0.328	-0.062
10	0.473	*0.074*	0.178	*0.032*	0.469	*0.043*	0.295	0.004
11	0.436	*0.076*	0.110	*0.031*	0.293	*0.044*	0.326	0.143
12	0.644	*0.100*	0.295	*0.043*	0.616	*0.059*	0.349	0.027
13	0.921	*0.096*	0.670	*0.043*	0.741	*0.062*	0.250	0.180
14	0.896	*0.107*	0.336	*0.045*	0.437	*0.062*	0.559	0.459
15	0.663	*0.091*	0.111	*0.039*	0.521	*0.051*	0.552	0.142
16	0.537	*0.089*	0.170	*0.036*	0.078	*0.046*	0.367	0.459
17	0.692	*0.096*	0.226	*0.044*	0.649	*0.055*	0.466	0.043
18	0.488	*0.095*	0.236	*0.028*	0.687	*0.043*	0.252	-0.198
19	0.730	*0.090*	0.202	*0.038*	0.435	*0.050*	0.529	0.295
20	0.560	*0.085*	0.467	*0.030*	0.495	*0.042*	0.094	0.065
21	0.365	*0.093*	0.052	*0.028*	0.429	*0.041*	0.312	-0.064
22	0.369	*0.100*	0.065	*0.023*	0.515	*0.037*	0.304	-0.146
23	0.282	*0.114*	-0.002	*0.019*	0.692	*0.039*	0.284	-0.410

Note: Standard errors in italics.

Estimating Spillover

Table B-2 shows our results for estimates of spillover for both the voluntary and default CPP rate. Spillover is estimated in the same way as the peak load reductions are estimated for the RCT method defined earlier in Appendix B, except that instead of using the peak hour electricity consumption on event days, we estimate the load reduction during the three baseline calculation periods (discussed in Section 2.1): (1) pre-peak hours on the event day; (2) peak hours on baseline days; and (3) pre-peak hours on baseline days.

Table B-2. Estimates of Spillover

Default CPP

Event	**Pre-peak on Event Day**	*std error*	**Peak on 4 Baseline Days**	*std error*	**Pre-peak on 4 Baseline Days**	*std error*
1	0.033672	*0.042018*	0.146214	*0.044096*	0.072093	*0.031294*
2	-0.00776	*0.047336*	0.14467	*0.044096*	0.071218	*0.031294*
3	-0.0162	*0.0421*	0.143604	*0.044095*	0.070445	*0.031293*
4	0.094833	*0.038535*	0.141677	*0.044096*	0.069757	*0.031294*
5	0.079323	*0.043351*	0.141677	*0.044096*	0.069757	*0.031294*
6	-0.02979	*0.04803*	0.141677	*0.044096*	0.069757	*0.031294*
7	0.026604	*0.042526*	0.141329	*0.044095*	0.069289	*0.031294*
8	0.043156	*0.040119*	0.141329	*0.044095*	0.069289	*0.031294*
9	0.037492	*0.03606*	0.149287	*0.044095*	0.073377	*0.031294*
10	-0.02076	*0.039487*	0.149287	*0.044095*	0.073377	*0.031294*
11	0.037981	*0.037806*	0.149287	*0.044095*	0.073377	*0.031294*
12	0.0921	*0.057073*	0.149322	*0.044096*	0.072952	*0.031295*
13	-0.00589	*0.062555*	0.148704	*0.044096*	0.072461	*0.031295*
14	0.063611	*0.061471*	0.148704	*0.044096*	0.072461	*0.031295*
15	0.062383	*0.048749*	0.149585	*0.044097*	0.073272	*0.031296*
16	0.098345	*0.047079*	0.14964	*0.044098*	0.073183	*0.031296*
17	0.032663	*0.053902*	0.149406	*0.044098*	0.073068	*0.031296*
18	0.075307	*0.047177*	0.149656	*0.044098*	0.073195	*0.031296*
19	0.040037	*0.051146*	0.149656	*0.044098*	0.073195	*0.031296*
20	0.094324	*0.046677*	0.149656	*0.044098*	0.073195	*0.031296*
21	0.062431	*0.045019*	0.149985	*0.044098*	0.07334	*0.031297*
22	0.074516	*0.04804*	0.150176	*0.044098*	0.07344	*0.031297*
23	0.070702	*0.046799*	0.150276	*0.044098*	0.073529	*0.0313*

Voluntary CPP

Event	**Pre-peak on Event Day**	*std error*	**Peak on 4 Baseline Days**	*std error*	**Pre-peak on 4 Baseline Days**	*std error*
1	0.134158	*0.064043*	0.176702	*0.041426*	0.046441	*0.036183*
2	0.214715	*0.074901*	0.172021	*0.040759*	0.051181	*0.035269*
3	0.034575	*0.063566*	0.214227	*0.041343*	0.05388	*0.035645*
4	0.128164	*0.062818*	0.146898	*0.042217*	-0.01254	*0.03578*
5	0.111345	*0.069381*	0.146946	*0.042217*	-0.0125	*0.035781*
6	0.127892	*0.074275*	0.146956	*0.042217*	-0.01246	*0.035781*
7	0.128428	*0.067336*	0.141927	*0.042439*	-0.00448	*0.035813*
8	0.15737	*0.063129*	0.141817	*0.04244*	-0.00457	*0.035813*

9	0.098356	*0.062172*	0.176392	*0.045818*	0.106355	*0.040666*
10	0.07614	*0.061404*	0.176306	*0.045816*	0.106343	*0.040666*
11	0.146282	*0.065142*	0.176342	*0.045816*	0.106352	*0.040666*
12	0.065146	*0.094225*	0.22134	*0.053364*	0.07783	*0.044966*
13	0.103142	*0.10074*	0.253097	*0.051009*	0.053687	*0.043245*
14	0.109463	*0.096634*	0.253412	*0.051006*	0.053893	*0.043245*
15	0.15733	*0.077519*	0.206065	*0.05183*	0.092684	*0.046299*
16	0.067053	*0.076289*	0.23054	*0.051652*	0.063966	*0.043681*
17	0.015589	*0.085869*	0.209381	*0.050227*	0.056247	*0.041429*
18	0.104298	*0.072886*	0.209114	*0.052666*	0.056312	*0.042351*
19	0.111275	*0.080648*	0.208769	*0.052664*	0.056109	*0.042352*
20	0.204919	*0.076656*	0.208948	*0.052663*	0.05623	*0.04235*
21	0.199939	*0.082709*	0.226079	*0.057189*	0.087935	*0.044842*
22	0.139831	*0.091734*	0.250886	*0.062124*	0.11713	*0.049002*
23	0.166574	*0.087081*	0.323639	*0.072752*	0.144512	*0.063097*

Note: Standard errors in italics.

Estimating Bias from Spillover

Table B-3 shows our results for calculations of bias from spillover and the portion of total 4-in-5 bias explained by spillover for both the voluntary and default CPP rate.

In order to quantify how spillover contributes to total bias, we define a “bias from spillover” that calculates how spillover during the three baseline calculation periods affects the 4-in-5 baseline, and therefore biases the peak load reduction estimates. Bias from spillover is calculated as described in Section 2.5.

Table B-3. Calculations of Bias from Spillover and the Portion of Total 4-in-5 Bias Explained by Spillover

Default CPP

Event	**Total Bias of 4-in-5**	**Bias as % of Savings**	**Bias from Spillover**	**Portion of Total Bias Explained by Spillover**
1	0.31474	13%	0.107793	0.34
2	0.023779	94%	0.065696	
3	0.19359	32%	0.056962	0.29
4	0.237835	17%	0.166754	0.70
5	0.18306	47%	0.151243	0.83
6	0.004146	99%	0.042133	
7	-0.20118	174%	0.098643	
8	0.065743	80%	0.115196	
9	0.223365	14%	0.113401	0.51
10	0.234589	-8%	0.055152	0.24
11	0.194278	6%	0.11389	0.59
12	0.38168	7%	0.16847	0.44
13	-0.04422	111%	0.070356	
14	0.377213	20%	0.139854	0.37

15	0.364462	-12%	0.138696	0.38
16	0.376443	2%	0.174802	0.46
17	0.509039	11%	0.109001	0.21
18	0.197819	41%	0.151769	0.77
19	0.492464	-17%	0.116498	0.24
20	0.032631	91%	0.170786	
21	0.300939	16%	0.139076	0.46
22	0.20095	27%	0.151252	0.75
23	0.121871	20%	0.147449	

Voluntary CPP

Event	**Bias of 4-in-5adj**	**Bias as % of Savings**	**Bias from Spillover**	**Portion of Total Bias Explained by Spillover**
1	0.444635	49%	0.26442	0.59
2	0.345601	67%	0.335556	0.97
3	0.261286	58%	0.194922	0.75
4	0.382545	47%	0.287599	0.75
5	0.350741	60%	0.270787	0.77
6	0.333788	64%	0.287307	0.86
7	-0.01102	102%	0.274839	
8	0.163447	76%	0.30376	
9	0.327924	35%	0.168393	0.51
10	0.294681	38%	0.146102	0.50
11	0.326117	25%	0.216272	0.66
12	0.348997	46%	0.208655	0.60
13	0.250263	73%	0.302552	
14	0.559442	38%	0.308982	0.55
15	0.551929	17%	0.270711	0.49
16	0.366811	32%	0.233627	0.64
17	0.46599	33%	0.168722	0.36
18	0.251812	48%	0.257099	
19	0.528527	28%	0.263935	0.50
20	0.093608	83%	0.357638	
21	0.31235	14%	0.338083	
22	0.303874	18%	0.273587	0.90
23	0.284333	-1%	0.345701	

References

[1] Institute for Electric Innovation: **Utility-Scale Smart Meter Deployments: Building Block of the Evolving Power Grid**; 2014.

[2] FERC: **Assessment of Demand Response & Advanced Metering: Staff Report**. Washington, D.C.; 2016.

[3] Wang Y, Li L: **Critical peak electricity pricing for sustainable manufacturing: Modeling and case studies**. *Applied Energy* 2016, **175**:40-53.

[4] KEMA: **PJM Empirical Analysis of Demand Response Baseline Methods**: PJM Markets Implementation Committee; 2011.

[5] George S, Bode J, Berghman D: **2012 San Diego Gas & Electric Peak Time Rebate Baseline Evaluation**: San Diego Gas & Electric; 2013.

[6] Goldberg ML, Agnew GK: **Measurement and Verification for Demand Response - Prepared for the National Form on the National Action Plan on Demand Response: Measurement and Verification Working Group**; 2013.

[7] KEMA: **New York Independent System Operator Special Case Resource Baseline Study**; 2014.

[8] Potter JM, George SS, Jimenez LR: **SmartPricing Options Final Evaluation**: U.S. Department of Energy; 2014.

[9] Braitwait SD, Hansen DG, Armstrong DA: **2011 Impact Evaluation of San Diego Gas & Electric's Peak-Time Rebate Pilot Program**; 2012.

[10] Baylis P, Cappers P, Jin L, Spurlock A, Todd A: **Go for the Silver? Evidence from field studies quantifying the difference in evaluation results between "gold standard" randomized controlled trial methods versus quasi-experimental methods**. In: *ACEEE Summer Study on Energy Efficiency in Buildings: August 21-26, 2016 2016; Asilomar, CA*. ACEEE.

[11] George S, Perry M, Woehleke S: **2010 Load Impact Evaluation of San Diego Gas & Electric Company's Summer Saver Program**: San Diego Gas & Electric; 2011.

[12] Lalonde RJ: **Evaluating the econometric evaluations of training programs with experimental data**. *The American Economic Review* 1986:604-620.

[13] DellaVigna S: **Psychology and Economics: Evidence from the Field**. *Journal of Economic Literature* 2009, **47**(2):315-372.

[14] Brown J, Hossain T, Morgran J: **Shrouded attributes and information suppression: Evidence from the field**. *The Quarterly Journal of Economics* 2010, **125**(2):859-876.

[15] Lacetera N, Pope DG, Sydnor JR: **Heuristic thinking and limited attention in the car market**. *The American Economic Review* 2012, **102**(5):2206-2236.

[16] Acland D, Levy MR: **Naivete, projection bias, and habit formation in gym attendance**. *Management Science* 2015, **61**(1):146-160.

[17] Gabaix X, Laidson D: **Shrouded attributes, consumer myopia, and information suppression in competitive markets**. *The Quarterly Journal of Economics* 2005, **121**(2):505-540.

[18] Gabaix X, Laidson D, Moloche G, Weinberg S: **Costly information acquisition: Experimental analysis of a boundedly rational model**. *The American Economic Review* 2006, **96**(4):1043-1068.

[19] Rabin M, Thaler RH: **Anomalies: risk aversion**. *The Journal of Economic Perspectives* 2001, **15**(1):219-232.

[20] Borghans L, Duckworth AL, Heckman JJ, Ter Weel B: **The economics of psychology of personality traits**. *Journal of Human Resources* 2008, **43**(4):972-1059.

[21] Harrison GW, Rutstrom EE: **Risk aversion in the laboratory**. In: *Risk aversion in experiments.* Edited by Limited EGP; 2008: 41-196.

[22] Hartog J, Ferrer-i-Carbonell A, Jonker N: **Linking measured risk aversion to individual characteristics**. *Kyklos* 2002, **55**(1):3-26.

[23] Harrison GW, List JA, Towe C: **Naturally occurring preferences and exogenous laboratory experiments: A case study of risk aversion**. *Econometrica* 2007, **75**(2):433-458.

[24] Kim T, Lee D, Choi J, Spurlock CA, Sim A, Todd A *et al*: **Extracting baseline electricity usage using gradient tree boosting**. In: *International Conference on Big Data Intelligence and Computing (DataCom 2015): 2015; Chengdu, Sichuan, China*. IEEE.

[25] Cappers P, Scheer R: **American Recovery and Reinvestment Act of 2009: Final Report on Customer Acceptance Retention, and Response to Time-Based Rates from Consumer Behavior Studies**. Berkeley, CA; 2016.

[26] Luo Y, Spindler M: **L2-Boosting for Economic Applications**. *The American Economic Review* 2017, **107**(5):270-273.

[27] Kozbur D: **Testing-based forward model selection**. *The American Economic Review* 2017, **107**(5):266-269.

[28] Varian HR: **Big data: New tricks for econometrics**. *Journal of Economic Perspectives* 2014, **28**(2):3-27.

[29] Athey S, Imbens GW: **Machine learning methods for estimating heterogeneous causal effects**. *stat* 2015, **1050**(5).

[30] Athey S, Imbens GW, Pham T, Wager S: **Estimating average treatment effects: Supplementary analyses and remaining challenges**. *The American Economic Review* 2017, **107**(5):278-281.

[31] Chernozhukov V, Chetverikov D, Dmirer M, Duflo E, Hansen C, Newey W: **Double/Debiased/Neyman Machine Learning of Treatment Effects**. *The American Economic Review* 2017, **107**(5):261-265.

[32] Imbens GW, Wooldridge JM: **Recent developments in the econometrics of program evaluation**. *Journal of Economic Literature* 2009, **47**(1):5-86.

[33] Thayer D, Brummer W, Smith BA, Aslin R, Cook J: **Is behavioral energy efficiency and demand response really better together?** In: *2016 ACEEE Summer Study on Energy Efficiency in Buildings: 2016; Asilomar, CA*. ACEEE.