When You Can’t Go for the Gold:
What is the best way to evaluate a non-RCT demand response program?

Olivia Patterson, Seth Wayland and Katherine Randazzo, Opinion Dynamics, Oakland, CA

ABSTRACT

As smart thermostats become more attractive to program administrators, evaluators are striving to find the most appropriate methodologies to assess both demand response and energy impacts associated with these devices. In the last handful of years, many utilities have been offering smart thermostats to customers as pilot initiatives, providing evaluators with an array of opportunities to test the accuracy, bias and costs of assessing impacts using differing methodological approaches, as well as the intersection of both energy efficiency and demand response impact approaches. Results from these pilot efforts help to inform future program design (e.g., how and when to call demand response events, and for whom; the types of cycling strategies to employ; how behavior can affect energy impacts; and internal and external validity of estimates).

In this paper, we provide estimated demand response impacts for a residential smart thermostat pilot using two different methodologies – a randomized controlled trial (or experimental design) and an quasi-experimental design. The results provide insights as to how to best design demand response program event protocols, balancing the program implementation goal of achieving as much demand reduction as possible with the evaluation goal of ensuring high degrees of accuracy and limited bias.

Introduction

Randomized control trials (RCTs) are widely regarded as the gold standard of impact evaluation, yet, program administrators cannot always implement them due to legal, financial, or other considerations. These problems are particularly acute for demand response (DR) programs, which are typically offered due to capacity constraints. Excluding customers to create a control group could result in a failure to reduce enough demand from the grid during peak periods. Consequently, evaluators must frequently estimate program impacts without a control group. It is therefore critical to understand how the choice of baseline estimation methods affects demand reduction estimates, and which estimation method is best to use under different circumstances.

This paper uses data from a 2015 evaluation of a smart thermostat pilot to evaluate two approaches to estimating demand response program impacts. The pilot program included 1,440 residential participants. Although participants opted into the program, we randomly assigned participants to treatment and control status during individual demand response events. Because we deployed the pilot events as an RCT, we compared the results from a non-RCT approach to the unbiased savings estimates from the RCT.

Specifically, this paper compares the performance of a common impact evaluation method: Mahalanobis Distance Day Matching using linear fixed effects regression (LFER) to an RCT design. In addition to evaluating the performance of this analytic approach, this paper pays attention to the relative performance under a very common condition: when event days are much hotter than most or all comparison days.

This has important implications not only for evaluators tasked with calculating impacts for programs that do not use RCTs, but also for utilities and program implementers who must decide whether delivering a program as an RCT is worth the cost. It also provides insight into how much error, from both bias and accuracy, we can expect from these non-RCT-based methods. A better understanding of the performance of non-RCT evaluations methods will help ensure that evaluators' results and recommendations truly help "make ambitious reductions real".

Bring Your Own Smart Thermostat Pilot

Opinion Dynamics evaluated a residential smart thermostat pilot, which offered customers an incentive for participating in DR events via a web-addressable smart thermostat. The objective of the pilot was to identify whether there were sufficient demand response, energy efficiency and bill savings associated with the pilot as an alternative to the utility's existing load control switch technology. Overall, the evaluation assessed kWh, kW, load shift and bill savings to inform future program design.

The Pilot employed a Bring Your Own Thermostat (BYOT) delivery model, which uses vendor-driven marketing approaches to achieve customer enrollment needs. The Pilot offered over ten different thermostat devices, via multiple vendors, who used their own marketing strategies (often varied and diverse) to enroll customers into the Pilot. The Pilot offered an incentive of $100 for a new thermostat, as well as an incentive for customers with an existing thermostat to participate in DR events.

Overall, a little under 1,500 customers enrolled in the Pilot, which were substantially fewer participants than anticipated. In total, the program administrator called nine events during the summer event season, which ran from May through September 2015, with a maximum of 40 hours of event participation per participant. Table 1 provides a list of the DR events called during the Pilot period.

Table 1. Demand Response Events Called during Pilot Period

Event Number	Event Date	Event Time	Temperature
1	July 7, 2015	2–6 PM	75°F
2	July 10, 2015	2–6 PM	77°F
3	July 28, 2015	2–6 PM	86°F
4	July 29, 2015	2–6 PM	88°F
5	August 18, 2015	2–6 PM	78°F
6	September 1, 2015	2–6 PM	81°F
7	September 2, 2015	2–6 PM	81°F
8	September 16, 2015	2–6 PM	82°F
9	September 18, 2015	2–6 PM	75°F

The evaluation used Advanced Metering Infrastructure (AMI) data at a sub-hourly level for the summer event period. We excluded from the analysis the 15% of participants without AMI data.

Study Methodology – It's All About the Counterfactual…

Best practice suggests delivering programs using an experimental design, using random assignment to treatment and control groups (an RCT). This approach simplifies the analysis and increases confidence in the estimate (Rubin 1974, Shadish, 2002). However, random assignment reduces the possible demand reductions for the program because some participants will not participate in the event (e.g., the control group). RCT designs therefore, can be expensive and time-consuming, are sometimes difficult to implement well, and are often impossible in real-world contexts. In addition, while a well-designed and implemented RCT will inspire confidence in the program impacts, the results may not apply more generally to future populations.

Quasi-experimental designs provide an alternative approach to estimating impacts, but they also have drawbacks. Using a quasi-experimental approach means that we must identify comparison group customers that are as similar as possible to program participants in aspects related to energy consumption and program participation. Identifying these customers is difficult, and can often produce biases that are difficult to quantify. When selecting a comparison group, we must match or control for a host of threats to validity, including History, Selection, Maturation, Statistical Regression to the Mean, Testing, and Instrumentation (Campbell and Stanley 1963).

Clearly, the ideal situation for estimating DR event impacts is to randomly assign customers to either a treatment or a control group, which would mean that we use the same (event) days as points of comparison. This approach eliminates bias from self-selection by participants and from different comparison days producing valid, unbiased, estimates. This paper, by having an RCT, allows our team to test quasi-experimental designs against RCT results. This effort is consistent with recent energy efficiency evaluation efforts that have sought to quantify the size and direction of biases associated with using an experimental or quasi-experimental design (Spurlock 2017, Smith 2015).

For this study, we estimated ex post demand reduction impacts for a smart thermostat pilot using two distinct approaches. The first is an experimental design (RCT) using a difference-in-difference (DID) model for customers randomly assigned to treatment and control groups for each event. The second approach uses a quasi-experimental design, Mahalanobis distance day matching with an LFER, to estimate demand reductions. We describe the two research designs below Table 2.

Table 2. Overview of Research Design, Matching and Modeling Approaches Employed

Research Design	Matching Approach	Modeling Approach	Control
Experimental	Random Assignment	Difference-in-Difference (DID)	Event Day
Quasi-Experimental	Mahalanobis Distance Day Matching	LFER	Similar Day

Experimental Design

Opinion Dynamics developed an RCT to estimate DR impacts of events called during the peak summer period of 2015. In this approach, for each event called, one half of the enrolled customers served as a control group and did not participate in a DR load control event. For the next event, those control group customers received the event and those who received the first event served as control. This is a trade-off between maximizing total DR impact and being able to provide accurate, unbiased estimates of per-participant DR impacts. Table 3 displays the treatment and control assignment, with one-quarter of the participants assigned to one of four groups. In each event, two groups were treatment and two control, which was varied by event. We randomized into four groups so that we could estimate consecutive day impacts using a crossover design.

Table 3. Group Assignment for Each Event

	Group			
Event	A	B	C	D
1	Treatment	Control	Treatment	Control
2	Treatment	Control	Control	Treatment
3	Control	Treatment	Control	Treatment
4	Control	Treatment	Treatment	Control
5	Treatment	Control	Treatment	Control
6	Treatment	Control	Control	Treatment
7	Control	Treatment	Control	Treatment
8	Control	Treatment	Treatment	Control
9	Treatment	Control	Treatment	Control

The RCT design allows for a simple DID modeling approach for the demand impact analysis. Because customers were randomly assigned into treatment and control groups, the average control group usage during the event is a valid counterfactual baseline for what the treatment group's usage would have been had they not been selected to participate. As a result, we can calculate savings simply by subtracting average hourly treatment group usage from control group usage during each event hour.

Quasi-Experimental Design

Given that we had estimated impacts using an experimental design, we saw an opportunity to leverage the same data to estimate impacts using quasi-experimental approach to identify the size and direction of the bias associated with this approach, and to identify any tradeoffs with either evaluation approach. In a quasi-experimental design for DR events, we match similar days using a reference load day approach. Reference load days provide information about what participants' consumption would have been on event days if the event were not called. For this reason, it is important for reference days to be as similar as possible to the event days. To select reference days that are most similar to event days, we used Mahalanobis Distance Day Matching to select reference days with weather profiles that are closest to the event days.

Not all days make good proxies for an event day. Cool days, when air conditioning is not used, for instance are not comparable to the hottest days, when events are called. The reference day values simulate the load the customer would have had if the DR event had not been called, also known as the counterfactual or baseline. Using this matching technique, we selected non-event days that best matched the hourly profile of each event day.

Mahalanobis Distance Day Matching minimizes the difference between the event and non-event day temperatures at each hour, correcting for the measured variation in temperature at that hour and the correlation of temperature between hours. To estimate baseline usage correctly, the matched days must cover the range of temperatures experienced on event days (black) and non-event days (gray). Figure 1 provides event day and non-event day temperatures prior to matching, while Figure 2 provides Mahalanobis Distance Day Matching event and matched non-event day temperatures, respectively.

Figure 1. Event Day and Non-Event Day Temperatures before Matching

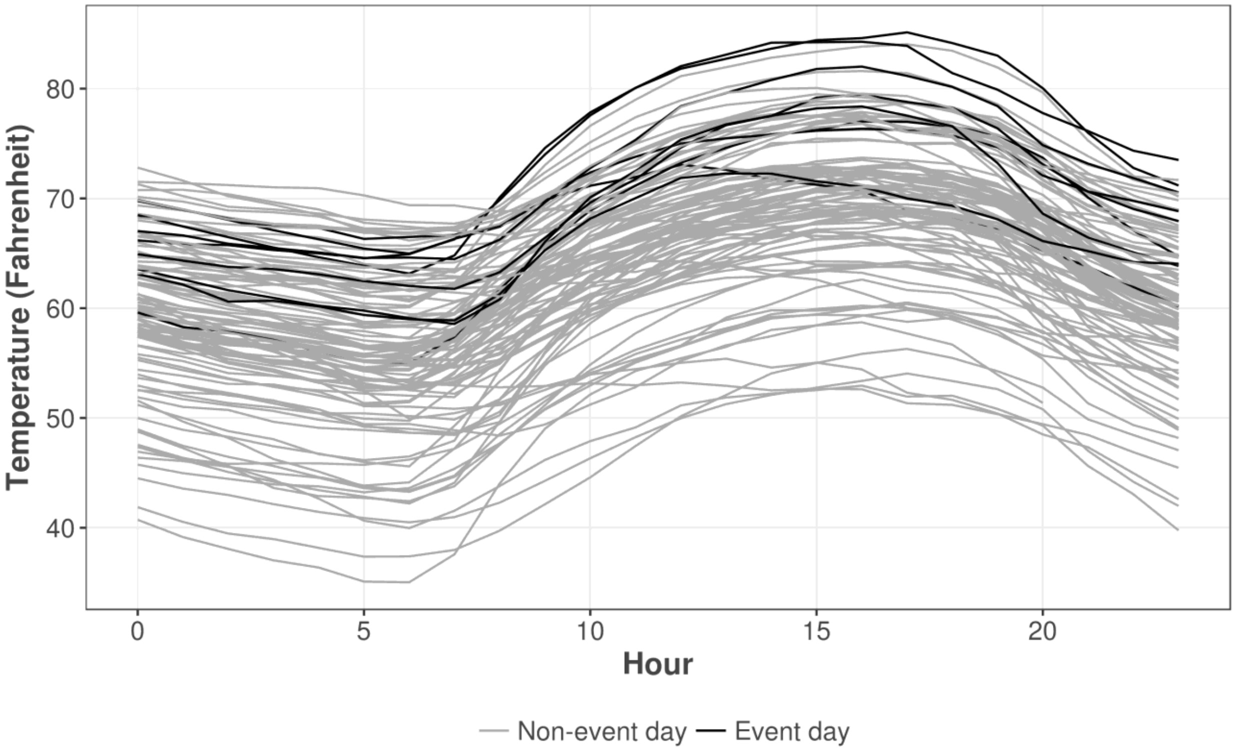
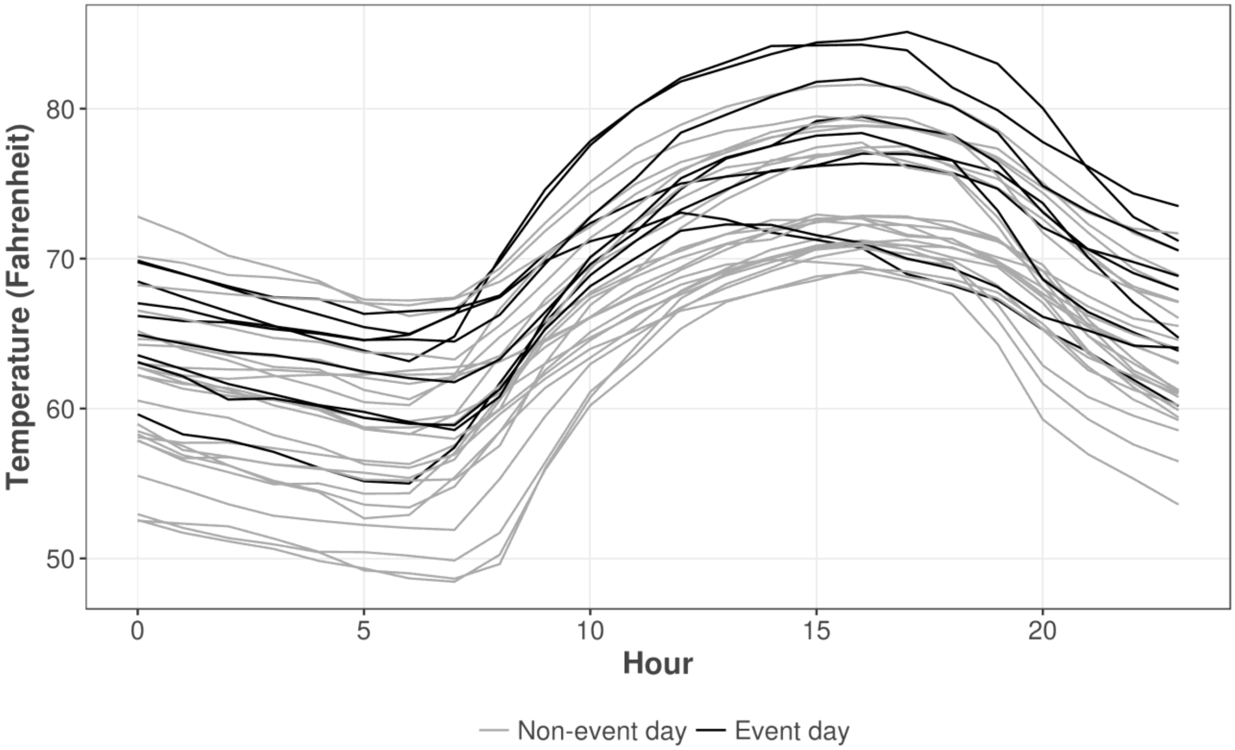
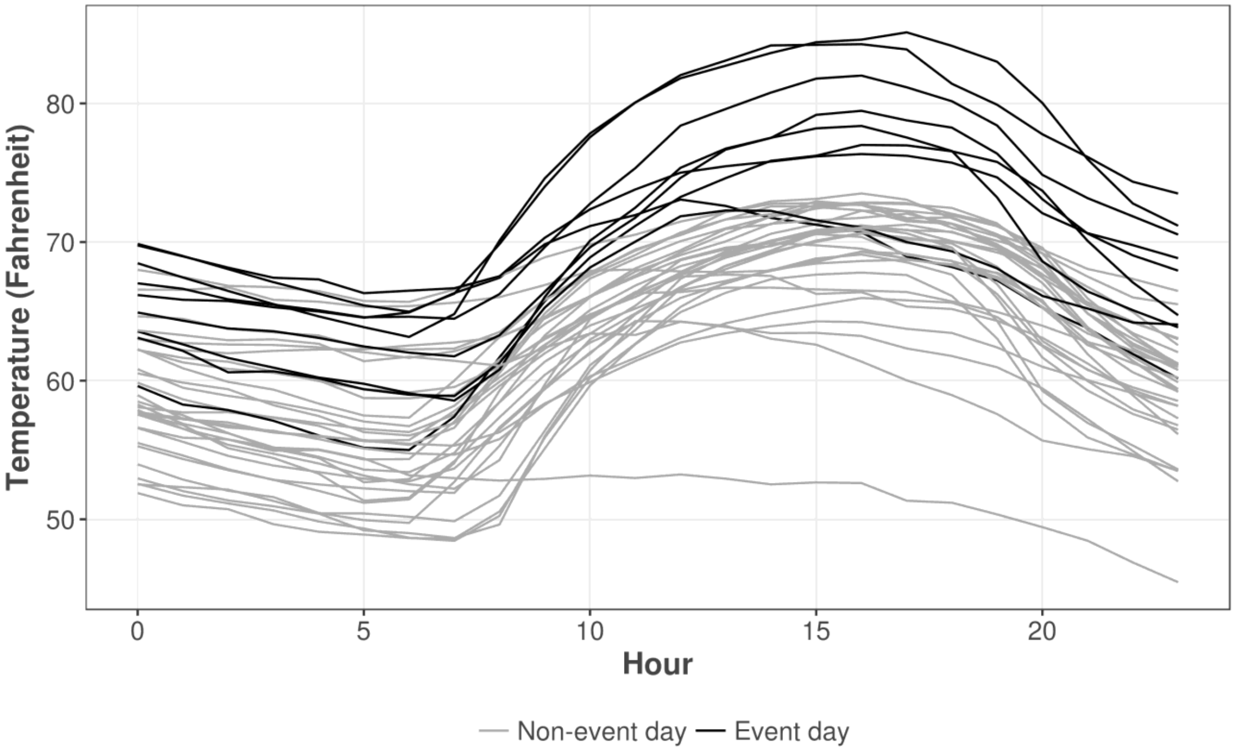


Figure 2. Event Day and Non-Event Day Temperatures after Mahalanobis Matching



We also conducted analyses to derive insights as to the effects of model estimates and uncertainty associated with well-matched versus poorly-matched reference days. To achieve this, we removed the best comparison days from the data and ran our analysis without those days. Figure 3 provides the event and matched non-event day temperatures using less than ideal candidates for reference load days.

Figure 3. Event Day and Non-Event Day Temperatures after Mahalanobis Matching -- Poorly Matched



For the quasi-experimental analysis, we estimated demand response impacts using a LFER modeling approach. This model produces demand reduction estimates associated with each day and hour, as well as an average demand reduction for all events called during the event season. It also accounts for all time-invariant, household-level factors affecting energy use without explicitly measuring those often-immeasurable factors and entering them explicitly in the models. These fixed-effects are contained in a household-specific intercept.

We selected the LFER model specification to best predict reference load during event days. The selected model incorporates weather variables (e.g., cooling degree hours (CDH)), as weather is one of the major predictors of energy consumption for AC use. The model also includes the hour of the day, as time of day is highly predictive of usage. We specified a broad range of models to ensure that the chosen model estimates baseline usage during events as accurately as possible. Equation 1 in the Appendix shows the final model that we used to estimate hourly demand reductions.

Study Results – Methods Matter…

Below we provide results from our study. Importantly, these demand reduction results are specific to a particular service territory with unique energy consumption, climate zone and housing stock, and should not be applied to other jurisdictions.

Table 4 provides the results of both research designs for ease of comparison. Both well-matched and poorly-matched quasi-experimental demand impact estimates are biased low when compared to the

RCT results, though the well-matched weather day's model is less biased. This result is unsurprising given that the RCT does a better job at identifying the best control, e.g., the actual weather, humidity, and day, on the day of the event. Further, those with poorly-matched days (e.g., with cooler temperatures) produce the lowest per thermostat savings (0.10 kW per hour). An experimental design produces the least biased, most accurate impact results, suggesting that program administrators should strongly consider incorporating these designs into their smart thermostat program evaluations.

Table 4. Summary of Ex Post Impact Results from Pilot, by Research Design

Research Design	**Matching Approach**	**Modeling Approach**	**Reference kW**	**Per Thermostat Kw Demand Reduction Savings**	**Standard Error**	**95% CI Daily Savings**	
						Lower	**Upper**
Experimental	Random Assignment	DID	1.88	0.45	0.01	0.42	0.48
Quasi-Experimental	Well-Matched Mahalanobis Distance Day Matching	LFER	1.75	0.37	0.01	0.35	0.40
	Poorly-Matched Mahalanobis Distance Day Matching	DID	1.48	0.10	0.01	0.08	0.13

Figure 4 and Figure 5 provide a visual depiction of the demand reductions that occurred during an average event for well-matched and poorly-matched estimates. The black line reflects actual observed energy consumption on an hourly basis, and the gray line reflects the estimated baseline consumption for a non-event day. The gray bar is the event period. As can be seen, the difference between the black and gray lines in Figure 5 is smaller than in Figure 4.

Figure 4. Average 2015 Summer Ex Post Demand Response Event Impacts (Well Matched)

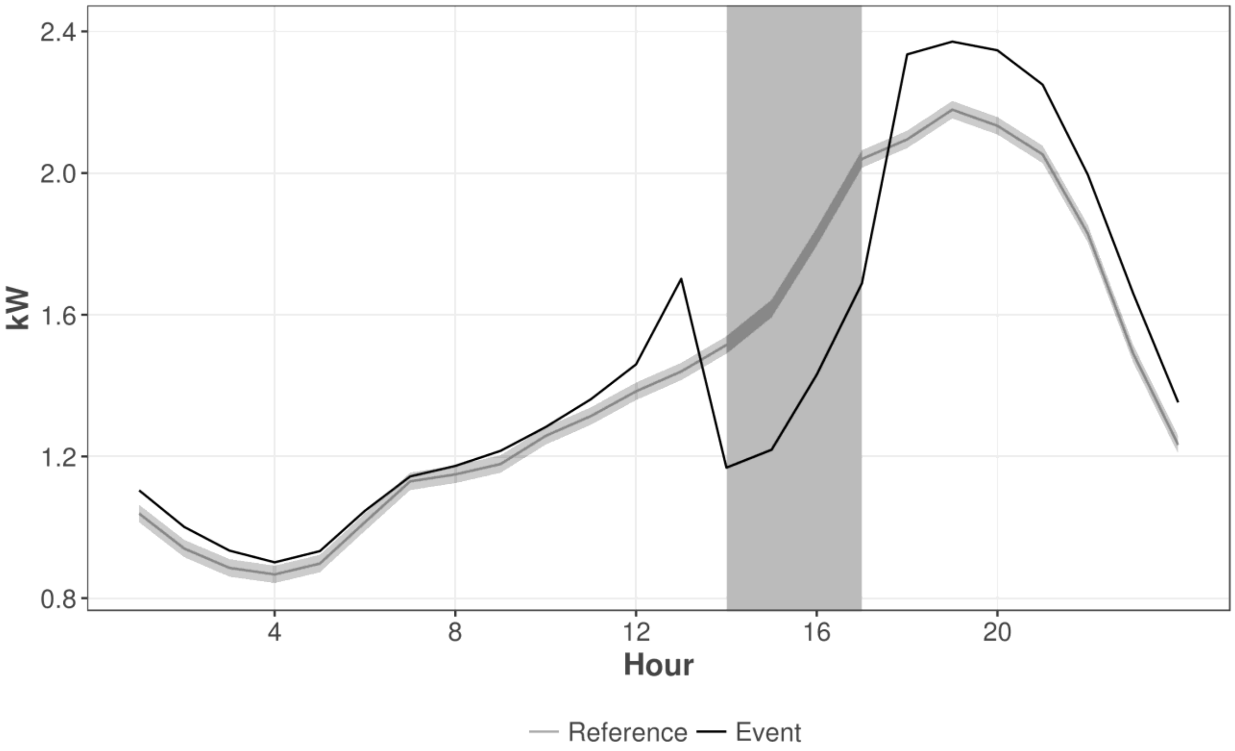
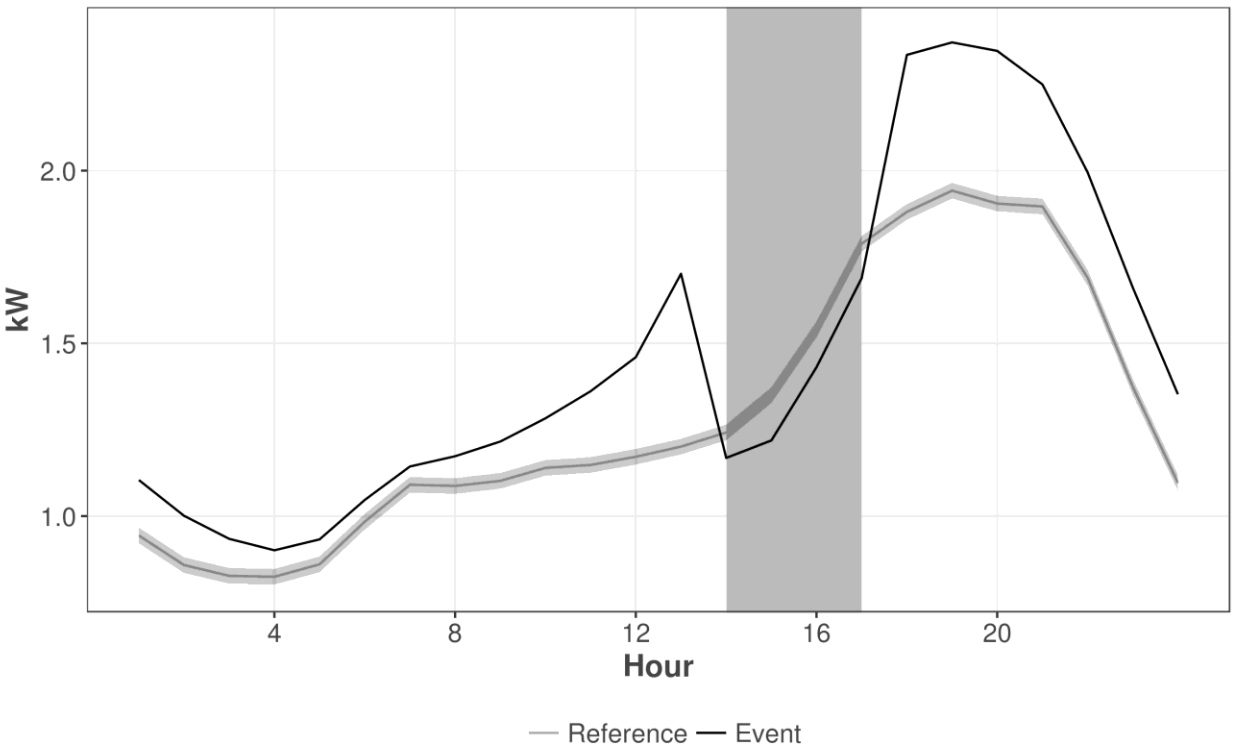


Figure 5. Average 2015 Summer Ex Post Demand Response Event Impacts (Poorly Matched)



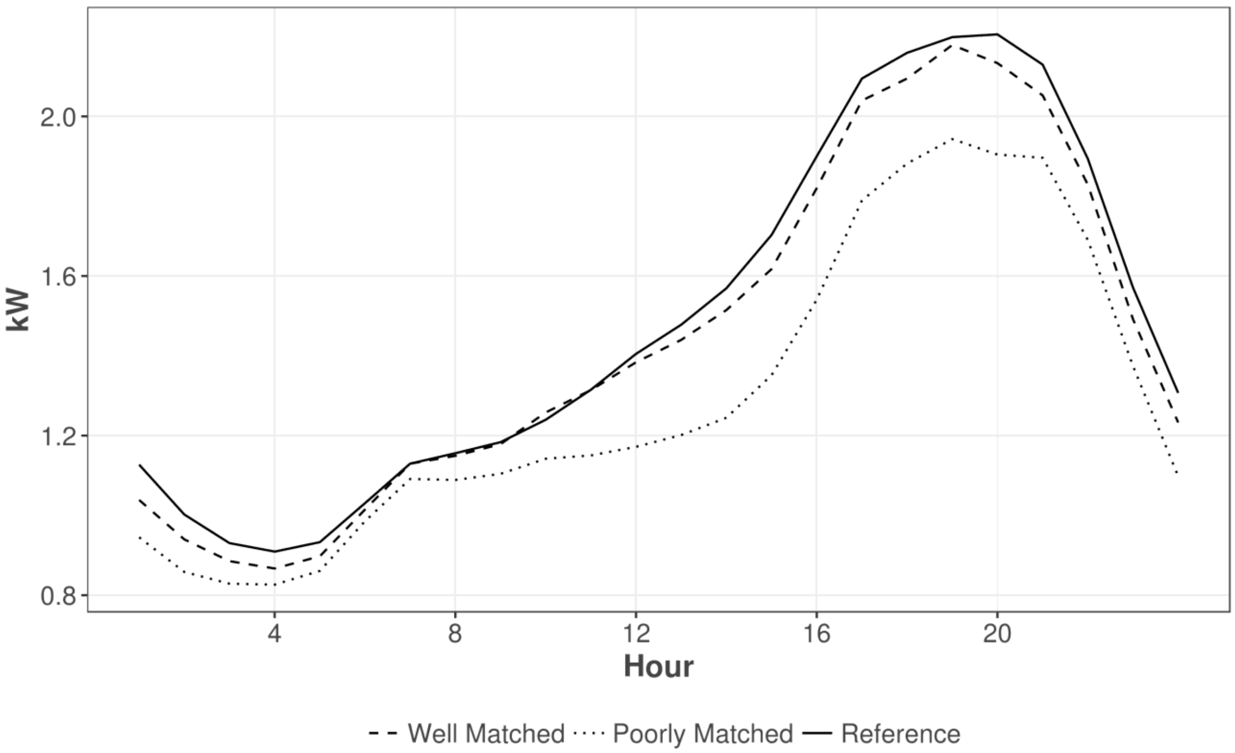
We tested a range of models before choosing a final specification. The primary method for evaluating the validity of a LFER model is to compare actual, logged runtime to the runtime predicted by the model. When actual and modeled usage are similar, especially on non-event days with weather similar to event days, it shows that the model is effectively estimating the reference load.

The figures below compare actual reference runtime from the RCT control on event days to the event day reference predicted by models. The actual runtime is higher than that from both models, but the well-matched quasi-experimental model does a much better job of approximating reference load than the model based on poorly matched days.

In addition to predictive power, we also considered adjusted R-squared and Akaike's Information Criterion (AIC)[1] in model selection. We chose models that maximized adjusted R-squared and minimized AIC. We used bootstrapped variance estimation to adjust confidence intervals for heteroscedasticity and autocorrelation.

[1] AIC balances predictive power and model parsimony, and thus helps guard against overfitting. For more, see: Palin and Haugh. "Eliminating the Guesswork: The Information Theoretic Approach to Model Selection." 2007 International Energy Program Evaluation Conference, Chicago, IL.

Figure 6. Average RCT Reference vs Modeled Reference Event Day Usage



Conclusion

This analysis confirms that it is possible to design an effective RCT for a thermostat DR program, and that the RCT experimental design produces the least biased, most accurate, demand reduction results, without substantial overhead due to design or implementation. As a result, we recommend that program implementers utilize an RCT approach for event days, where feasible.

This approach can be effective when the program is being offered as a pilot or when reducing the total number of participants who receive a load control signal during an event will not jeopardize the load reduction required for peak system emergencies. It can also be offered as a way to reduce event participation requirements, possibly increasing volunteer participation or mitigating opt-outs during events. Finally, if program administrators offer smart thermostats from multiple vendors or thermostat device types, or utilize distinct cycling strategies, an RCT can help to increase the accuracy of demand reduction estimates with smaller population sizes.

This analysis also suggests that when using a quasi-experimental design, poorly-matched (or cooler temperature) weather days may produce an underestimation of overall program impacts. This is most likely due to events falling on some of the hottest days of the season, which leaves mostly cooler days with lower demand as comparison days. The regression model attempts to correct for differences in weather and other effects, but because there are many days hotter than the event days, the models underestimate baseline consumption.

It is clear from these results that poorly matched comparison days with much lower temperatures lead to more bias and reduced impact estimates. This bias shows that carefully selecting comparison days is very important to accurately predicting reference load. One solution is to call test events during average weather days to mitigate this issue. In addition, evaluators should report model validation statistics to demonstrate how well the weather days matched event days, and what types of interpretations are appropriate based on the modeled results.

When quasi-experimental results are biased low compared to the RCT results, evaluators underestimate both impacts and cost-effectiveness. This could potentially lead to cancelling or reducing the size of a program that was performing well. The opportunities to claim savings (if estimates are biased low), or risks to the grid (if estimates are biased high and the program is in a capacity constrained area) are real issues that program planners and administrators continue to face with load management efforts.

Results from these pilot efforts can also help to inform future program design (e.g., how and when to call demand response events, and for whom; the types of cycling strategies to employ; how behavior can affect energy impacts; and internal and external validity of estimates). RCT designs may also better support understanding participant engagement with events (such as opt-out behavior or thermostat usage), that can contribute to optimizing program messaging and targeting for future event seasons because it can also establish a counterfactual for what similar participants would have done on an event day without an event being called.

When attempting to balance an RCT design with a quasi-experimental design, consider the total number of participants in the program, the overall contribution to the load, and develop control groups that are appropriately scaled to estimate impacts accurately. We encourage evaluators to work with program implementers prior to any event season to assess whether or not it is appropriate and feasible to develop an RCT experimental design in support of estimating demand reductions for a pilot or program. This requires additional upfront work to develop the random assignment of customers, and work with implementers to maintain fidelity to the design during the event season.

Appendix

Below we provide our estimating equation for the ex post regression model as well as model validation results.

Estimating Equation

Equation 1: Ex Post Regression Model

$$
\begin{aligned}
kw_{it} = \mu_0 + \alpha_i + \beta_{event} \cdot Event + \sum_{r=1}^{4} \beta_{region\ r} \cdot Region_r + \sum_{r=1}^{4} \beta_{event\ region\ r} \cdot Event \cdot Region_r \\
+ \beta_{CDH} \cdot CDH_t + \sum_{r=1}^{4} \beta_{region\ r\ cdh\ t} \cdot Region_r \cdot CDH_t + \sum_{h=1}^{23} \beta_{hour\ h} \cdot Hour_h \\
+ \sum_{h=1}^{23} \beta_{event\ hour\ h} \cdot Event \cdot Hour_h + \sum_{h=1}^{23} \beta_{event\ hour\ h} \cdot Event \cdot Hour_h \cdot CDH_t \\
+ \sum_{m=6}^{9} \beta_{month\ m} \cdot Month_m + \sum_{m=7}^{9} \sum_{h=1}^{23} \beta_{month\ hour\ mh} \cdot Month_m \cdot Hour_h + u_i + \varepsilon_{it}
\end{aligned}
$$

Where:

kw_{it} = Hourly energy consumption – load in hour *t* for customer *i* (kWh/hr)

μ_0 = Overall mean energy usage

α_i = Participant-specific deviation from mean energy usage

u_i = Participant-specific error

ε_{it}= Observation-specific error

$Event$ = Indicator variable for event day for those participants in the treatment group

$Hour$ = Set of 23 indicator variables for the hours of the day

$Month$ = Set of 4 indicator variables for the months of the program (May–Sept)

$Region$ = Set of 4 indicator variables for region

CDH = Cooling degree hours

References

California Public Utilities Commission Energy Division. "Attachment A: Load Impact Estimation for Demand Response: Protocols and Regulatory Guidance." April 2008. http://www.calmac.org/events/FinalDecision_AttachementA.pdf

Campbell and Stanley, (1963). Experimental and Quasi-Experimental Designs for Research. New York: Houghton Mifflin Company.

Rubin, D. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, Vol. 56, 5, (1974) 488-701.

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.

Smith, Brian A. and Josh Shellenberg 2015. “Comparison of Methods for Estimating Energy Savings from Home Energy Reports.” *Behavior, Energy and Climate Change Conference.* http://beccconference.org/wp-content/uploads/2015/10/presentation_smith_brianarthur1.pdf

Spurlock, A., Cappers, P.L., Jin, L., Todd, A., and Patrick Baylis, “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.” *ACEEE Conference 2017.* http://aceee.org/files/proceedings/2016/data/papers/2_363.pdf