No Customer Left Behind: Innovation in Evaluation to Expand the Reach of Energy Efficiency Programs

Kayla Banta, Sarah Monohon, Charles Hanks, Evergreen Economics
David Tripamer, Bonneville Power Administration
Thomas Elzinga, Central Electric Cooperative
Jake Harrison, Tacoma Power

ABSTRACT

High-volume, low-savings interventions such as home energy reports have been proven to deliver cost-effective energy reductions at scale. The low cost to distribute these interventions, paired with demonstrated, consistent energy savings, offer utilities an avenue for easy wins in helping their customers save energy and money.

However, current evaluation requirements for behavioral programs, such as deploying interventions through a randomized control trial (RCT), may prevent utilities from serving customers as fully as they could. Utilities face a difficult choice: 1. Design an RCT that withholds treatment from a random subset of eligible customers to enable accurate measurement of savings impacts or 2. Provide the treatment to all eligible customers to achieve greater savings but forfeit the ability to claim those savings (because no control group exists for measurement purposes). To make matters more difficult, small utilities do not always have the sample size necessary to detect small savings from home energy reports, even with an RCT. Alternative evaluation approaches are necessary to help these small utilities expand program access while maintaining sufficient evaluation rigor to claim those savings.

This paper presents lessons learned from analysis of a behavioral program deployed to a small utility's customer base with no control group. We use a quasi-experimental evaluation method that combines site-level baseline modeling, customer segmentation via k-means clustering on model-derived features, and nearest-neighbor matching within clusters to construct a comparison group from non-participants. We also validate our methods against a utility that offered the same product within a classic RCT experimental design.

This alternative evaluation method could unlock universal deployment of proven low-cost interventions where an RCT is not practical and hourly usage data are available. The lessons learned will be presented in terms of tradeoffs for maximizing energy savings and achieving a reasonable level of certainty around realized savings, demonstrating how the goals of multiple perspectives (program administrators and evaluators) may be balanced with flexibility and innovation.

Introduction

For utilities seeking to meet energy efficiency targets, behavioral energy programs such as home energy reports programs are an attractive proposition due to their high volume, low cost, and proven results. For evaluators, the randomized control trial (RCT) is the gold standard for estimating savings of these programs. In an RCT design, eligible customers are randomly assigned to treatment and control groups, allowing evaluators to measure the energy savings achieved by the treatment group that are attributable to the program, above and beyond any natural changes in energy consumption observed in the control group.

According to the Regional Technical Forum, home energy reports are personalized communications to utility customers that compare their energy to similar households and provide tailored conservation tips (2022).[1] They have been a staple of energy efficiency portfolios since the mid-2000s. Programs typically deliver savings of 1–2 percent of annual household consumption, with a meta-analysis of recent evaluations finding an average of 88 kWh per year in electric savings per household (Munson et al. 2022). Because the savings are small relative to normal variation in household energy use, the opt-out RCT became the standard evaluation design. The US Department of Energy's Uniform Methods Project (UMP) formalized this approach, identifying the RCT as the recommended method for behavioral program evaluation.

However, this rigorous experimental design comes at a cost. First, detecting relatively small energy savings requires large analysis samples to achieve sufficient statistical precision. The UMP notes that samples in the thousands or tens of thousands of households are often needed (Stewart and Todd 2020), placing opt-out RCT designs effectively out of reach for smaller utilities. Second, the design requires that a random subset of eligible customers be withheld from the intervention for the duration of the program. This is a loss for those customers, who could otherwise be saving energy, and a loss to the utility, which can only claim savings observed within the treatment group. These challenges raise the question of whether every behavioral program should be required to use an RCT, or whether alternative methods can provide sufficient evidence under the right conditions.

Prior efforts to use quasi-experimental methods have failed to replicate RCT estimates for home energy reports programs. Schellenberg et al. (2015) compared propensity score matching, Bayesian structured time series, and regression tree models against a large-scale RCT benchmark. Propensity score matching overstated savings by 89 percent, while the time series and machine learning models were highly unstable. The authors noted that none of the methods tested used advanced metering infrastructure (AMI) data, and called for further research using hourly interval data.

This paper responds to that gap, comparing two pathways to detect savings from a behavioral program implemented without an RCT: a pooled monthly difference-in-differences (DiD) with a matched comparison group, and a cluster-matched DiD leveraging site-level hourly baseline models. Using data from Oregon's Central Electric Cooperative's (CEC) Behavioral Home Energy Report program, we test whether the more sophisticated approach can produce reliable savings estimates where simpler methods fall short.

Program Background

Bonneville Power Administration (BPA) markets wholesale electrical power to a network of consumer-owned customer utilities across Washington, Oregon, Idaho, and Montana. CEC is a small electric utility serving approximately 35,000 residential customers in the Pacific Northwest (FindEnergy 2024). BPA works with customer utilities like CEC to acquire energy conservation as a resource resulting from approved programs (BPA 2025a).

[1] The Regional Technical Forum (RTF) is a stakeholder body operating under the Northwest Power and Conservation Council that develops standardized energy savings estimates for energy efficiency measures and programs in the Pacific Northwest. RTF-approved values are used by Bonneville Power Administration and its customer utilities to claim savings toward regional conservation targets.

In addition to its energy conservation portfolio, BPA also offers an Open Q pilot program. The Open Q pilot program provides a pathway for utilities to pilot energy conservation programs that do not fit within BPA's pre-approved programs (BPA 2025b). This paper focuses on a Behavioral Home Energy Report program that CEC submitted to BPA's Open Q pilot program.

In December 2023, CEC began sending emails with behavioral patterns and energy tips to approximately 26,000 residential customers for whom it had an email address and who had not opted out of communications. The remaining 9,000 customers did not receive these energy emails either because CEC lacked an email address on file or because they had actively opted out. Similar to a traditional Home Energy Reports Program, eligible customers received monthly email reports providing personalized feedback on their energy consumption relative to similar neighbors. CEC was using a proven product, with savings that had been demonstrated with an RCT in other regions. They did not set aside a control group for an RCT so they could serve all of their customers. BPA and CEC sought a reliable method for quantifying and validating the program's savings.

Methodology

Data Collection

CEC provided Evergreen Economics with hourly AMI data for all eligible treated customers and all ineligible untreated customers, spanning August 2022 through December 2024, along with monthly billing records covering November 2021 through December 2024. All consumption data were joined with hourly temperature readings from the nearest National Oceanic and Atmospheric Administration (NOAA) weather station, with timestamps standardized to UTC. After cleaning consumption and weather data, our final sample consisted of 21,179 participants and 6,850 non-participants.

Creating Matched Comparison Groups

When a randomized control group is not available, the UMP acknowledges that alternative methodologies should be considered, particularly for small utilities facing practical constraints (Stewart and Todd 2020). In these cases, developing a matched comparison group and conducting DiD analysis is a well-established quasi-experimental approach.

Matching reduces pre-existing differences between participants and non-participants, while the DiD framework accounts for natural consumption trends affecting both groups, isolating the program effect on energy use. This approach assumes that all customers, participants and non-participants, experience similar natural changes in energy consumption over time. A DiD approach is only as accurate as the comparison group it uses.

Figure 1 shows the challenge of constructing a reliable comparison group for the participant population. Participants consistently used more electricity in the pre-period compared to non-participants, with a gap between the hours of 4 p.m. and 9 p.m. In a true RCT implementation, we would expect that the blue and red lines would overlap one another perfectly, with a near zero-kWh difference in usage at each hour. This gap means that without careful matching, a DiD model risks attributing some of baseline usage differences to the treatment effect.

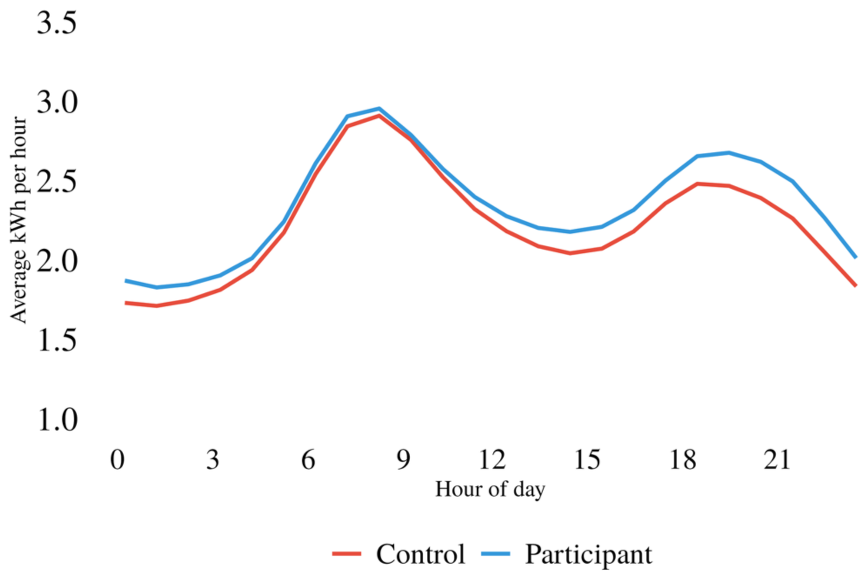

Figure 1. Participant and non-participant hourly load shapes in pre-period

Two more observations about the usage data complicate this approach:

1. Both participants and non-participants consumed approximately 1,000 kWh less in the post-period relative to the pre-period. This trend is 10 times the magnitude of the *ex-ante* savings signal, 100 kWh. A natural change of this size would obscure the program effect in a simple pre-post analysis with no comparison group. Because this trend affects both groups, a well-matched comparison group can remove it from the program impact estimate. However, its large magnitude means that even small differences between matched pairs can produce a biased savings estimate.
2. Year-to-year site-level consumption changes showed high individual variation in both groups, with roughly 15 percent of customers increasing usage by more than 10 percent and 30 percent of customers decreasing usage by more than 10 percent. The *ex-ante* savings of 100 kWh per year is modest relative to the natural variation occurring across the customer population regardless of program participation.

While a regression model can control for these pre-existing differences in energy usage across the two groups, our concern is that the groups may have different drivers of energy consumption. If participants have larger homes, more occupants, or more plug loads, then their natural changes in energy consumption from pre- to post-treatment may differ from the non-participants. Our goal with matching is to find the non-participant whose consumption trajectory most closely mirrors what the participant would have done absent the program. We describe two methods below that differ in how much information they use to create these matched pairs.

Method 1: Monthly Pooled Difference-in-Differences

The pooled monthly DiD model is a standard approach for evaluating residential energy efficiency programs. We present this method as a benchmark to show the limitations of simple matching on monthly usage. Below we describe this comparison group matching and savings estimation method.

We used one-to-one nearest-neighbor matching to construct a comparison group of non-participants. We calculated mean monthly kWh consumption during the pre-period (August 2022 – July 2023) for each participant and non-participant, then selected the single best match for each participant based on minimized root mean square error (RMSE) across the 12 pre-period months. Matches were required to be within 75 percent and 150 percent of the participant's baseline consumption. This method resulted in 18,321 matched pairs out of a pool of 18,329 participants and 6,708 non-participants; matching was done with replacement so that multiple participants could be matched to the same non-participant. The average pre-period absolute difference between matched participants and controls was 11.9 kWh/month or 0.75 percent of average monthly consumption. The standardized mean difference on pre-period monthly load was 0.0013, indicating excellent balance between the two groups.

We estimated program effects using a fixed effects DiD specification:

$$kWh_{i,t} \sim \beta 0 + \beta_1 Part_i + \beta_2 \mathrm{Post_t} + \beta_3 (\mathrm{Part_i} \times \mathrm{Post_t}) + \beta_4 HDD_{i,t} + \beta_5 CDD_{i,t} + \varepsilon_{i,t}$$

Where:

$kWh_{i,t}$ = Electricity consumption of customer *i* at end of month *t*
$Part_i$ = Indicator variable equal to 1 if participant, 0 if non-participant
$Post_t$ = Indicator variable equal to 1 for the post-period, 0 for pre-period
$HDD_{i,t}$ = Heating degree days at site of customer *i* at the end of month *t*
$CDD_{i,t}$ = Cooling degree days at site of customer *i* at the end of month *t*
β values = Parameters to be estimated
$\varepsilon_{i,t}$ = Random error term

The coefficient β_3 on the $\mathrm{Part_i} \times \mathrm{Post_t}$ interaction represents the estimated monthly energy savings attributable to the program.

Method 2: OpenDSM Cluster-Matched Difference-in-Differences

Method 2 uses site-level baseline models and k-means clustering to construct a comparison group. Each site's model coefficients form a behavioral signature for how and when a customer uses electricity, and how much electricity they use. Method 2 follows three steps: 1) fit a separate baseline model for each site, 2) cluster all customers by their model-derived behavioral signatures, and 3) match each participant to a non-participant within the same cluster. We then estimate savings by differencing out the natural change in usage observed in matched non-participants.

OpenDSM Hourly Baseline Modeling

We fit a separate baseline model for each customer using OpenDSM, an open-source hourly energy modeling package. Each model is trained on 12 months of pre-period hourly consumption data using elastic net regression.[2] The model specification is:

$$kWh_{i,t} = \beta_0 + \sum_k \beta_k T_k(temp_t) + \sum_j \gamma_j C_{j,t} + \sum_{j,k} \delta_{jk} \left(C_{j,t} \times T_k(temp_t) \right) + \varepsilon_{i,t}$$

Where:

$kWh_{i,t}$ = Electricity consumption of customer *i* at hour *t*
β_0 = Site-specific intercept
$T_k(temp_t)$ = Piecewise linear temperature response within temperature bin k
$C_{j,t}$ = Indicator for temporal cluster *j* on day *t*, where clusters group days with similar load shapes
$C_{j,t} \times T_k(temp_t)$ = Interaction between temporal cluster and temperature bin
β, γ, δ = Site-specific parameters estimated via elastic net regression
$\varepsilon_{i,t}$ = Error term

Once the model is fit, we extract its coefficients as behavioral features for clustering and use the model to predict post-period consumption absent the program.

K-Means Clustering

A single model fit on the full population is weakened by the wide variation in electric usage. The model tries to explain many different behaviors at once with a single set of parameters. For example, some homes are more sensitive to hot days than others and use more electricity to cool their homes. A single population-level model must assign a single number (the cooling slope) representing the average home's response to a hot day despite a wide range of household behaviors. The more diverse the population in terms of cooling response, the less that single number represents any individual household, and therefore the model is less able to detect a modest program signal amidst all the noise.

In Method 1 we matched participants to non-participants based on average monthly consumption in the pre-period. While this ensures matched pairs used similar amounts of electricity overall, it did not account for how that electricity was used. Two homes can have identical average monthly consumption even if one has a high, stable baseline and the other has a low baseline with high electric heating loads that drive up the average. Customers who look similar in level but differ in their underlying usage drivers may diverge in the post-period for

[2] Elastic net regression shrinks irrelevant coefficients toward zero, which reduces overfitting the model (OpenDSM 2026). Standard regression models fit to a single site's historical data risk overfitting: learning the noise and idiosyncrasies of that meter's history rather than the underlying relationship between weather and consumption. An overfitted model produces predictions that track the training data closely but perform poorly on new data. Elastic net regression addresses this by applying a penalty that shrinks small or redundant coefficients toward zero, keeping only the temperature and load-shape relationships. This produces more reliable post-period predictions than a standard regression would at the individual site level (OpenDSM 2026).

reasons unrelated to the program, creating bias in the DiD estimate. This is important for a behavioral program: home energy reports work by prompting customers to change how and when they use electricity. A participant and non-participant who use the same amount of electricity and share similar usage habits are more likely to have followed similar usage trajectories absent the program.

We clustered customers on features extracted from site-level baseline models. These features describe how each customer's usage responds to temperature and time-of-use patterns, making them a better representation of usage behavior than raw aggregate usage values alone. We then used k-means clustering to assign each customer to the subpopulation whose average usage profile is closest to their own. We selected three clusters based on within-cluster variance and interpretability of the resulting groups. The goal of clustering is not to find perfect customer segments. Clustering customers reduces the burden on the nearest-neighbor matching algorithm, ensuring that each participant is compared to controls with similar usage patterns.

Figure 2 summarizes the behavioral differences across the three clusters. Each cell shows the cluster mean for that feature, with darker shading indicating higher values. Cluster 3 has the highest-use, most behaviorally consistent meters, averaging 3.46 kWh/hour with the strongest heating sensitivity (0.69) and the highest month-to-month consistency (0.50). Cluster 1 has the lowest average load but by far tends to have extreme usage events (kurtosis = 13.34) and the lowest month-to-month consistency (0.33).[3]

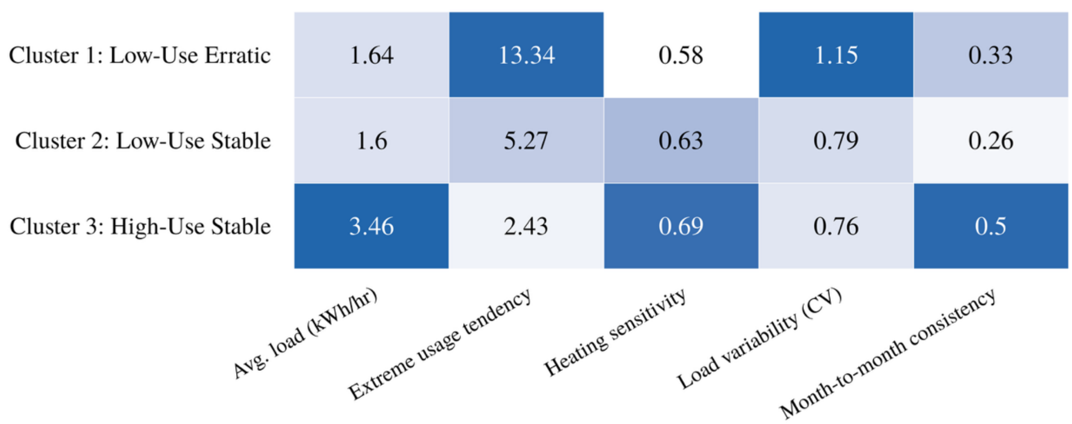

Figure 2. Behavioral differences across clusters

We summarize the characteristics of each cluster below:

1. **Cluster 1: Low-Use, Erratic (n = 4,529).** These customers have low mean consumption, high variability, frequent extreme spikes, and unpredictable month-to-month usage.
2. **Cluster 2: Low-Use, Stable (n = 8,816).** These customers have similar mean consumption to Cluster 1 but lower variability and more moderate usage, consistent with full-time occupants in smaller or more efficient homes.

[3] Kurtosis measures the tendency of a distribution to produce extreme values relative to its overall spread. In this context, a kurtosis of 13.34 in Cluster 1 reflects frequent large consumption spikes against an otherwise low baseline. For more information, see DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2(3)*, 292–307. https://www.columbia.edu/~ld208/psymeth97.pdf

3. **Cluster 3: High-Use, Stable (n = 9,239).** These customers have the highest mean consumption, strongest heating sensitivity (0.69), and most consistent month-to-month usage consistent with full-time owner-occupants in larger homes.

Table 1 shows the distribution of participants and non-participants across each of the three clusters:

Table 1. Distribution of CEC customers per cluster

	Participants		Non-participants	
Cluster	n	Percent	n	Percent
Cluster 1	3,583	20%	946	20%
Cluster 2	7,094	40%	1,722	36%
Cluster 3	7,107	40%	2,132	44%
Total	17,784	100%	4,800	100%

Matching

After assigning each meter to a cluster, we identified for each participant the single most similar non-participant within the same cluster, measured across four standardized variables that describe each meter's pre-period consumption:

- Average hourly consumption
- Coefficient of variation
- Skewness of consumption
- Number of distinct daily load patterns

We computed match distance as a weighted Euclidean distance across these four dimensions. Because overall usage is the strongest indicator of a comparable customer, average hourly consumption received a weight three times larger than the other matching variables. We matched with replacement, allowing multiple participants to share the same control meter. After matching, we applied a post-match quality filter, retaining only pairs where the average monthly pre-period consumption difference between participant and control was less than 500 kWh/month. This filter removed 75 matched pairs, or 0.4 percent of the sample, that had no adequate comparison in the non-participant pool. The final matched sample included 17,709 pairs, representing 84 percent of eligible participants (n = 21,179).

Figure 3 shows the average hourly load shapes for matched participants and controls in the pre-period. Matching reduced the mean hourly consumption difference between participants and controls from 0.122 kWh to 0.004 kWh, a 97 percent reduction in pre-period imbalance, and less than 0.2 percent of mean hourly consumption.

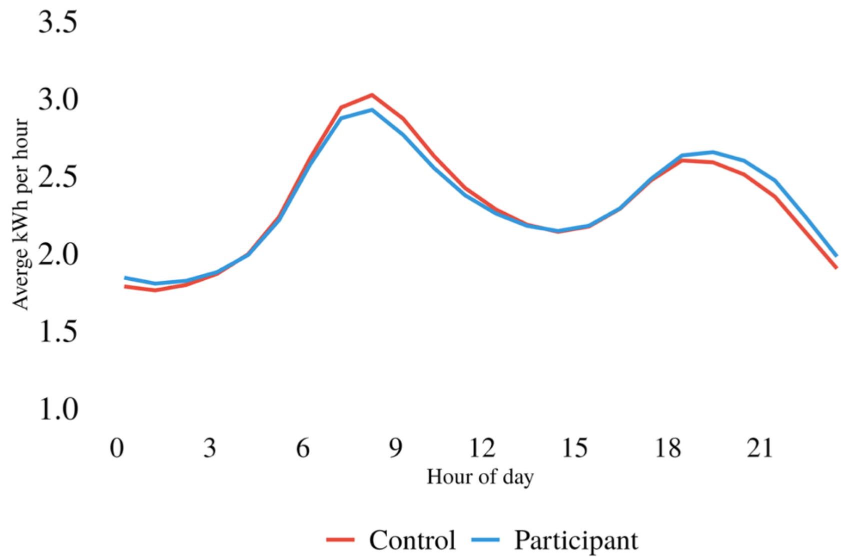

Figure 3. Matched participant and non-participant hourly load shapes in pre-period

Savings Estimation

We estimated program savings using three model specifications, each a different time resolution. The monthly and daily models apply the same fixed-effects DiD specification described in Method 1 to monthly and daily usage data respectively. We fit the monthly and daily models to check consistency of savings estimates and to inspect the participant coefficient, which captures the pre-period difference between matched participants and controls. If matched pairs are comparable, the participant coefficient should be small and statistically insignificant.

The hourly "collapsed DiD" uses site-level baseline models from OpenDSM to construct a weather-normalized counterfactual for each meter.

1. We fit the model specified above for each meter to predict hourly consumption.
2. We calculate the residual as predicted minus observed hourly usage to isolate the change in usage that is not explained by weather.
3. We compute the mean hourly residual for each meter separately in the pre- and post-periods:

$$\bar{\varepsilon}_{period} = mean(\hat{y}_{i,t} - y_{i,t})$$

Where $\hat{y}_{i,t}$ is the OpenDSM prediction at meter *i* at time *t* and $y_{i,t}$ is observed hourly kWh. This collapses each meter's hourly time series to two numbers: the average hourly residual (predicted minus observed) in the pre-period and in the post-period.

4. We calculate the DiD for each matched pair to remove the natural change that both participants and non-participants experience to isolate the treatment effect:

$$treatment\ effect = (\bar{\varepsilon}_{post,part} - \bar{\varepsilon}_{pre,part}) - (\bar{\varepsilon}_{post,nonpart} - \bar{\varepsilon}_{pre,nonpart})$$

The inner differences capture how each meter's average hourly residual changed from pre to post. A positive change means the meter used less than the baseline model predicted after the program began. The outer difference removes any pre-to-post change in residuals common to both participants and non-participants, isolating the change attributable to the program.

5. We estimate the average savings across all matched pairs by running a simple regression on the pair-level DiD values. This gives us the average savings estimate and the uncertainty around the estimate.

Results

Simple Pre-Post

Before presenting matched comparison group results, we first report the simple pre-post analysis without a comparison group. Participant weather-normalized consumption declined by approximately 298 kWh per year on average from the pre- to post-period, or 1.5 percent of their annual pre-period usage. Non-participant consumption also declined by approximately 159 kWh per year over the same period. The difference, roughly 139 kWh per year, is consistent with the pre-post with comparison group estimates below. Without a comparison group, a simple pre-post analysis overstates program savings by nearly half, attributing natural consumption trends to the program.

Pre-Post with Comparison Group

Table 2 presents savings estimates across both methods:

Table 2. Savings per evaluation method

Method		Time resolution	Point estimate (kWh/year)	P-value[4]	Significant (α = 0.10)
1	Fixed effects pooled DiD, matched on monthly usage	Monthly	65	0.342	No
2	Pooled DiD matched within cluster on OpenDSM-derived features	Monthly	124	0.080	Yes
		Daily	120	0.092	Yes
		Hourly	136	0.047	Yes

The monthly fixed effects model with participants matched on monthly usage alone produced an estimated annual savings of 65 kWh; however, this estimate is not statistically significant. When participants and non-participants have similar usage but different underlying drivers, they diverge in the post-period for reasons unrelated to the program, generating noise that overwhelms the small savings signal.

[4] Standard errors are clustered at the meter level for two reasons: serial autocorrelation of a meter's usage over time, and the repeated use of non-participants, where a single non-participant meter may serve as the matched control for multiple participant meters.

Method 2 resolves this by matching within clusters of behaviorally similar customers. The agreement across monthly, daily, and hourly specifications strengthens the findings that the signal is real rather than an artifact of any single method. Figure 4 shows the 90 percent confidence intervals of each method. A savings estimate of approximately 130 kWh/year represents about 0.7 percent of annual usage. This is a modest signal that must be detected against large natural variation in usage.

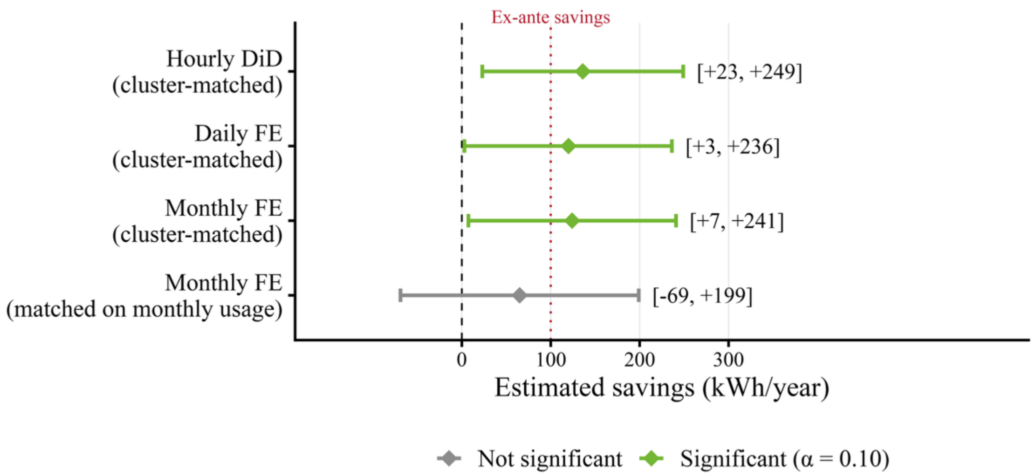

Figure 4: Savings estimates by matching method

If we compare the cost of evaluation relative to potential savings, the key question is whether regulators will accept quasi-experimental evidence. For example, let us assume that a utility has 25,000 eligible customers with email addresses, 5,000 with no email, estimated savings of 60 kWh per customer per year, and an electricity rate of $0.20 per kWh. In an RCT, roughly 15,000 customers would be treated, generating approximately $180,000 per year in avoided energy costs. After subtracting implementation and evaluation costs of approximately $36,000, the net program value is roughly $144,000 per year.[5] A BHER program that treats all 25,000 eligible customers (and about 5,000 customers with no email addresses being used as the comparison group) would generate approximately $300,000 per year. After the same implementation cost and a slightly higher evaluation cost of AMI-based analysis, net program value is approximately $254,000 per year. Allowing a BHER program to treat the large majority of eligible customers, rather than excluding half to a control group, far exceeds the increased cost of evaluation.

Method Validation

To assess whether Method 2 produces a savings estimate consistent with a true RCT, we obtained data from Tacoma Power, whose Home Energy Report program was implemented under an RCT design. We drew a sample of 14,644 treated customers, 6,776 control customers with email available, and 1,460 control customers without email. We then established a ground-truth savings estimate independent of any matching or clustering procedure given the random

[5] These costs are based on CEC's actual implementation contract and the incremental cost of adding one more utility-program-year to a regional home energy reports impact evaluation, which is standard practice in partnership with BPA.

assignment of treatment and control groups. Next, we tested whether our quasi-experimental method produced that same estimate when applied to a comparison group like that of the CEC evaluation, where non-participants were defined by the absence of an email address on file rather than by random assignment.

From the RCT benchmark, comparing the treatment and control-with-email groups, we detected an annual savings estimate of 72 kWh ($p = 0.018$). Method 2 produced an annual savings estimate of 84 kWh ($p = 0.20$). The two estimates are directionally consistent and close in magnitude, though the pooled quasi-experimental estimate is not statistically significant, likely reflecting the smaller sample size and control reuse introduced by within-cluster matching. Restricting the comparison group to non-participants without email produced a point estimate close to the RCT ground truth.

Conclusion

As the evidence base for behavioral programs has matured, evaluation frameworks have an opportunity to evolve alongside these programs. Quasi-experimental methods can be applied to estimate savings for a program without a randomized control group. The practical implication is significant. For small utilities where withholding treatment from a large control group is neither economically rational nor administratively feasible, quasi-experimental design is not a second-best option — it is the appropriate one. A well-matched comparison group, constructed with care before the regression is ever run, is the foundation that determines whether a quasi-experimental estimate will hold up to scrutiny.

For low-cost, high-volume interventions with an established RCT evidence base, a marginal improvement in savings precision from implementing an RCT is unlikely to deliver more value than treating more customers. A program that treats 25,000 eligible customers generates approximately $300,000 in avoided energy costs. After costs for the implementation contract and an AMI-based evaluation, net program value is roughly $254,000 — compared to approximately $144,000 under an RCT design that treats only 15,000 customers and uses monthly interval data for the evaluation. The key difference is in how many customers have access to savings. The incremental evaluation cost of hourly versus monthly billing data intervals is far less impactful than the decision to move away from an RCT.

The hourly AMI-based method found statistically significant savings while the monthly billing method did not. That result is not a failure of one method. It is evidence that the granularity of the data drives precision. If higher precision is required for regulators to trust the evaluation results and allow the program to continue without an RCT, then the added cost of hourly data will be worth it.

This points toward a reframing of what evaluation requirements are meant to accomplish for proven behavioral programs such as home energy reports programs. Rather than asking "did we find the right number?", the more productive question is "does the evidence contradict what we expected?" That standard preserves accountability while allowing utilities to expand program access and direct evaluation resources toward questions that generate more program value — understanding which customers respond and why, and how program design can improve.

References

BPA (Bonneville Power Administration). 2025a. Energy Conservation Implementation Manual 2026–2027. bpa.gov/-/media/Aep/energy-efficiency/document-library/implementation-manual-2026-2027-final.pdf

———. 2025b. "Open Q Pilot — BHER Program" [Slide deck]. bpa.gov/-/media/Aep/energy-efficiency/special-promotions/open-q-program-overview.pdf

FindEnergy. 2024. "Central Electric Cooperative." Published May 28, 2024. findenergy.com/providers/central-electric-cooperative/

Munson, T., M. Galport, K. Quaye, and C. Sloan. 2022. "Behavioral Programs Come of Age: Analyzing the Savings from Recent Home Energy Report Program Studies." In *Proceedings of the 2022 ACEEE Summer Study on Energy Efficiency in Buildings*, 8-55–8-69. Washington, DC: ACEEE.

OpenDSM. 2026. Hourly Model Methodology. Linux Foundation Projects. Published February 18, 2026. opendsm.energy/documentation/eemeter/hourly_model/methodology/

Regional Technical Forum. 2022. *Guidance on Evaluated Energy Savings for Residential Behavior Home Energy Reports.* Portland, OR: Northwest Power and Conservation Council. nwcouncil.app.box.com/v/0220222ResBehaviorGuideDoc

Schellenberg, J., A. Savage, M. Blundell, and J. Cook. 2015. *Comparison of Methods for Estimating Energy Savings from Home Energy Reports.* CALMAC Study ID PGE0364.01. Prepared for Pacific Gas and Electric Company. https://www.calmac.org/publications/Comparison_of_Methods_for_Estimating_HER_Energy_Savings_11032015.pdf

Stewart, J., and A. Todd. 2020. "Residential Behavior Evaluation Protocol." In *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*, Chapter 17. U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Washington, DC. https://docs.nlr.gov/docs/fy21osti/77435.pdf

AI-assisted Tools Disclosure

AI-assisted tools used: Claude
Used for: Drafting text, code assistance
Extent: Light editing, citation research, generating code
Human review: The authors reviewed and verified all content, calculations, and citations.