CPUC GROUP A - IMPACT EVALUATION REPORT

Population-Based NMEC - Program Years 2019 - 2021

California Public Utilities Commission
CALMAC ID: CPU0365.01

Date: December 22, 2023

Information	Details
Project Sponsor	Lullit Getachew
Project Manager	Lorre Rosen
Telephone Number	(510) 891-0446
Mailing Address	155 Grand Ave. Suite 600 Oakland, CA 94612
Email Address	lullit.getachew@dnv.com, lorre.rosen@dnv.com
Report Location	https://pda.energydataweb.com/

Table of contents

List of figures

List of tables

1 EXECUTIVE SUMMARY

This summary presents DNV's evaluation of the Population-Based Normalized Metered Energy Consumption (population NMEC) Programs for program years (PY) 2019 to 2021. We conducted this evaluation as part of the California Public Utilities Commission (CPUC) Energy Division (ED) Evaluation, Measurement and Verification contract.

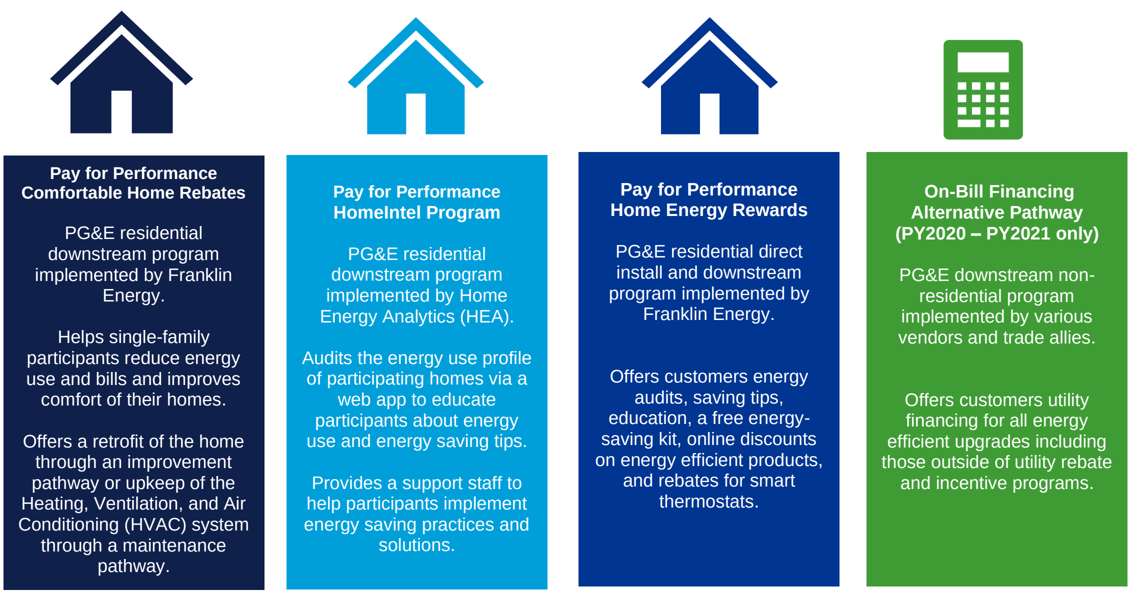
1.1 Study background

NMEC is a set of statistical tools and approaches that help determine the energy consumption impact of energy efficiency (EE) programs by comparing pre- and post-intervention meter data. While other EE programs claim final savings based on deemed[1] or calculated results, NMEC programs claim final savings based on measured impacts at the meter.

NMEC is an essential component of California's strategy to expand and deepen EE programs throughout the state, resulting from two bills (AB 802 and SB 350) encouraging utility companies to offer pay-for-performance (P4P) programs that use a data-driven approach to estimate EE impact. Our evaluation differs from other EE program evaluations that fall under this contract. For example, each NMEC program estimates its own performance following an overarching NMEC process (as defined by the NMEC Rulebook)[2] that must be evaluated.[3] We expect the current NMEC program evaluation process to change as the NMEC process matures further.

Figure 1-1 provides more information on the population NMEC programs selected for the PY2019 to PY2021 evaluations.[4] All of these programs were High Opportunity Project & Programs (HOPPs) programs that preceded the NMEC Rulebook.

Figure 1-1. Population-based NMEC programs for the PY2019 to PY2021 evaluation[5]



[1] Deemed refers to researched, vetted, and predictable savings for EE technologies and services with well-established properties. This contrasts with custom savings for EE technologies and services that require unique calculations and do not use predefined values.

[2] CPUC, "Rulebook for Programs and Projects Based on Normalized Metered Energy Consumption Version 2.0," 1/7/20, cpuc.ca.gov, https://www.cpuc.ca.gov/-/media/cpuc-website/files/legacyfiles/n/6442463694-nmec-rulebook2-0.pdf

[3] Independent evaluations determine the post-installation savings of non-NMEC EE programs based on an industry-vetted evaluation method instead of verifying savings produced by the embedded measurement and verification approach that program administrators use to pay implementers. Further, there are no required processes that independent evaluations of non-NMEC EE programs need to validate.

[4] Our evaluation does not include OBF PY2019 since it was evaluated in PY2018–2019 California Statewide On-Bill Financing Impact Evaluation – DRAFT CPUC Group B. March 7, 2023. Opinion Dynamics. https://pda.energydataweb.com/api/view/2780/CPUC%20OBF%20Evaluation%20Report_DRAFT_03092023_clean.pdf

[5] Downstream is a delivery mechanism that provides incentives and technologies directly to customers. Direct install programs provide energy-saving technologies or upgrades for no or low cost to participating customers through installation contractors.

1.2 Research objectives

Our research objectives in this evaluation were to:

- Verify gross and net savings[6] of residential population NMEC projects with claims in program years 2019, 2020, and 2021, and non-residential population NMEC projects with claims in program years 2020 and 2021.
- Assess the application of NMEC requirements outlined in the NMEC Rulebook.
- Determine the conditions for a successful evaluation of population NMEC programs.
- Characterize program participation.

1.3 Study approach

Gross savings. Program administrators (PAs) that claim savings based on population NMEC methods report initial savings in CEDARS[7] based on forecast or workpaper values. They true-up these initial claims based on metered results once 12 months of post-installation data become available. The sum of the initial and trued-up claims provides the final claimed savings for each program and year. The population NMEC programs we are evaluating followed this process for program years 2020 and 2021.[8]

In this instance, the PA added an additional step: it filed the trued-up savings to meet reporting deadlines but also undertook additional measurement and verification (M&V)[9] analysis that accounted for the effect of COVID-19 on P4P program savings. The PA did not complete the study by the true-up filing deadline, and the additional M&V results were not reflected in CEDARs claims. DNV based its evaluation on these revised PA M&V results, rather than on CEDARs claims.[10] We compared our evaluated results to the revised PA M&V savings to calculate the percent of claimed savings the programs achieved (evaluation realization rates). We also compared the evaluated savings to the final claimed savings to determine gross realization rates (GRRs).

Replication is an essential first step to developing an evaluation estimate of gross savings. For this evaluation, we attempted to replicate the PA approaches to determine savings for population NMEC projects with one year of post-installation data. We reviewed the statistical programming code and additional documents provided by the PA to replicate the methods used to calculate savings for P4P and OBF projects with installations from 2019 to 2021. As part of our replication, we ran input data provided by the PA through code we either translated or created based on a description of the methods provided. Where possible, we compared intermediate outputs to the corresponding data files provided. We also compared our results with the final results in M&V reports or outputs to determine the replicability of the results, a central tenet of the NMEC approach. While project calculations are replicated, they are also validated relative to best practices for consumption data analysis, as found in the Universal Method Protocol, chapter 8, and ASHRAE.[11]

There has only been one previous CPUC evaluation of a population NMEC program, based on 2018 and 2019 installations from PG&E's OBF program.[12] That evaluation covered all activities in the OBF program including population NMEC

[6] Gross savings are a measure of change in energy use due to energy efficiency programs, regardless of why customers participated. Net savings are changes in energy use attributable to a particular energy efficiency program and consider savings from participants who would not have purchased energy-efficient technologies without the program. Savings attributable to participants who would have purchased energy-efficient technologies with or without the program influence are excluded from net savings. These participants whom the program did not influence are considered free riders.

[7] California Energy Data And Reporting System.

[8] The PA did not true-up the PY2019 P4P program claims, and the final claimed savings are based on the initial claimed savings for this program year.

[9] M&V is the process of estimating the energy savings from an EE program intervention. It involves measuring the changes in the energy consumption of a particular technology or the whole premise due to this intervention.

[10] The PA did not revise its PY2020 OBF trued-up savings. Thus, DNV based its evaluation on the final claimed savings that reflect the true-ups reported in CEDARS for this program in PY2020.

[11] Agnew, Ken and Goldberg, Mimi. "Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol." National Renewable Energy Laboratory, November 2021. https://www.nrel.gov/docs/fy17osti/68564.pdf.

[12] See footnote 4.

(Alternative Pathway), site NMEC,[13] and Custom Projects.[14] The prior evaluation involved reviewing the NMEC methods used to determine savings and accepting the claimed savings of projects that could be verified. For this population NMEC evaluation, DNV went a step further and attempted to fully replicate the processes used to estimate P4P and OBF population NMEC project savings and validate the results. This evaluation sets a new benchmark of full replication and validation as the essential basis of a population NMEC program evaluation.

Program attribution and net savings. Population NMEC programs do not, in themselves, require novel evaluation methods to develop net-to-gross ratios (NTGRs). The OBF program did, however, require special attention because on-bill financing provides no-cost financing rather than a monetary incentive. We conducted web surveys with residential P4P and phone surveys with selected non-residential OBF participants to understand the percentage of installations that would have occurred without the programs.[15] The sample sizes collected for this purpose are shown in Table 1-1. The information from the surveys helped determine the NTGRs, which capture the degree to which customers would have installed program equipment or taken program-recommended actions without the program benefits. We used the estimated NTGRs to determine the amount of program savings that can be attributed to the programs (net savings).[16] NTGR estimates based on the sample sizes shown below satisfy the program-level 90/10 minimum confidence level and precision requirements.[17]

Table 1-1. Survey efforts and sample size summary

Surveys	Mode	Participant population	Population targeted	Completed surveys	Response rates
Residential P4P participant survey	Web	15,031	15,031	2,694	18%
Non-residential OBF participant survey	Phone	833	276	99	36%

Application of NMEC requirements. The evaluation team conducted a comprehensive review of the NMEC Rulebook, PG&E Advice Letters, program implementation plans (PIPs), and M&V reports to assess the application of NMEC requirements by the population NMEC programs under evaluation. In addition, we conducted interviews with PG&E program staff and various program implementers to gain insights into certain program aspects that were unavailable in the PIPs. To assess how the programs conformed with the NMEC guidelines outlined in the Rulebook, we examined how programs determined eligibility criteria, selected interventions and measures,[18] made incentive payments, and structured and planned M&V activities.

Evaluability of Population NMEC programs. Since this is one of the first population NMEC evaluations, we examined the foundations for successful evaluations of such programs. NMEC programs have embedded M&V, which PAs use to determine payable savings to implementers and claim program savings in CEDARS. The NMEC Rulebook requires these methods to be transparent and replicable and encourages programs to use publicly vetted approaches.

We use our current evaluation efforts to inform what is required for the successful validation of PA M&V savings by focusing on three areas. First, we address the challenges of replicability of PA M&V results to meet basic validation requirements. We then assess the PA M&V approach used to determine savings by examining if the methods used are sound and meet industry standards, if elements of the analysis reflect accepted criteria, and if the results are reasonable and statistically

[13] Site-NMEC involves meter-based savings analysis using pre- and post-period weather-normalized energy consumption for each participating site. In contrast, population NMEC involves pre-post analysis using weather-normalized energy consumption to determine average savings for similar participants and often uses a comparison group.

[14] See footnote 1.

[15] Population NMEC program attribution is based on data collected among P4P participants with installations from 2019 to 2021 and OBF participants with installations in 2020 and 2021.

[16] Gross savings are multiplied by the NTGR to arrive at net savings.

[17] Relative precision is a way to express how statistically confident we are that an answer is within a range. A 90/10 relative precision indicates that we are 90% confident that the true result falls within 10% on either side of the estimate.

[18] Interventions are energy-saving processes, while measures are energy-saving technologies.

significant. Third, we examine the challenges of determining how much of the claimed savings the program achieved (gross realization rates).

Program effectiveness. We examined the types of program participants that the population NMEC programs reached based on primary data we collected through web and phone surveys among participants. We collected information, including demographic-related, perceived program benefits, and satisfaction levels, to understand the characteristics and experiences of program participants. We used the information we collected and evaluated program savings to examine the effectiveness of the population NMEC programs.

1.4 Key findings

1.4.1 Program savings

1.4.1.1 Gross savings

This evaluation primarily focused on replicating and validating PA M&V savings estimates for the population NMEC programs. The PA developed the M&V savings estimates, in part, due to the effects of the pandemic on the accuracy of existing performance-based estimates. These M&V savings estimates superseded the claimed savings reported in CEDARS. The evaluation realization rates reflect the replication and validation work performed.

P4P. Table 1-2 summarizes the first year claimed, PA M&V, and evaluated (validated) savings. It also provides PA M&V savings relative to claimed (PA M&V realization rate) and evaluated relative to PA M&V savings (evaluation realization rate). The evaluation realization rates are close to 100% because DNV was able to replicate savings estimates for most participants and identified no major issues with how the PA estimated or assigned savings. The PA M&V process included baseline and reporting period exclusion conditions that identified conditions where performance-based savings could not be reasonably estimated and determined the type and timing of alternative assigned savings. Based on these exclusion conditions, the percentage of sites whose electric savings were entirely performance-based ranged from 50% to 60%, in line with billing analysis savings estimates that typically reflect 50% to 70% of participant data.[19] The subset of customers with zero assigned savings after they moved, opted out, added solar, or participated in other programs ranged from 11% to 15%. The assignment of zero savings is more conservative than in typical billing analysis practice and errs on the side of caution, which seems appropriate for NMEC programs. The results also indicate that two of the P4P programs achieved high percent savings in the last two program years, which met the NMEC Rulebook's requirement of 25% relative precision at 90% confidence.[20]

Table 1-2. Summary of P4P electric and gas gross savings, PY2019 to PY2021

Program	Program year	Electric sites	First year savings (kWh)							
			Claimed savings	PA M&V results	PA M&V realization rate	Evaluated savings	Evaluation realization rate	Gross realization rate	Percent savings	Relative precision
P4P - HomeIntel	2019	404	595,895	391,318	66%	391,318	100%	66%	7.5%	17%
	2020	818	643,945	561,852	87%	561,825	100%	87%	7.7%	15%
	2021	2,648	1,994,086	1,185,439	59%	1,185,409	100%	59%	6.7%	10%
P4P - Comfortable Home Rebates	2019	562	2,883	57,374	1990%	57,373	100%	1990%	1.1%	98%
	2020	2,053	638,753	861,411	135%	860,447	100%	135%	4.8%	14%
	2021	2,276	3,201,543	1,386,773	43%	1,383,443	100%	43%	7.2%	10%
	2019	484	38,085	102,994	270%	102,994	100%	270%	2.3%	53%

[19] The exclusion process for gas indicates an even higher percentage of performance-based savings.

[20] 90/25 indicates how confident we can be that an answer falls within a certain bound. It is a common criterion used in energy efficiency evaluation, requiring that the research achieves 90% confidence that an estimated metric, such as estimated savings, falls 25% on either side of the estimated value to provide a statistically valid outcome.

Program	Program year	Electric sites	First year savings (kWh)							
			Claimed savings	PA M&V results	PA M&V realization rate	Evaluated savings	Evaluation realization rate	Gross realization rate	Percent savings	Relative precision
P4P- Home Energy Rewards	2020	5,005	429,650	155,945	36%	155,597	100%	36%	0.5%	80%
	2021	1,779	673,671	108,948	16%	107,230	98%	16%	1.0%	112%
Program	**Program year**	**Gas sites**	**First Year Savings (therm)**							
			Claimed savings	**PA M&V results**	**PA M&V realization rate**	**Evaluated savings**	**Evaluation realization rate**	**Gross realization rate**	**Percent savings**	**Relative precision**
P4P - HomeIntel	2019	462	39,163	23,056	59%	23,025	100%	59%	6.1%	39%
	2020	789	79,960	14,479	18%	15,176	105%	19%	3.4%	37%
	2021	2,516	39,734	40,669	102%	40,720	100%	102%	3.5%	21%
P4P - Comfortable Home Rebates	2019	43	305	981	322%	994	101%	326%	3.9%	164%
	2020	1,346	24,052	36,531	152%	36,358	100%	151%	6.4%	18%
	2021	2,002	54,751	83,720	153%	83,818	100%	153%	9.1%	12%
P4P- Home Energy Rewards	2019	535	4,768	9,021	189%	8,847	98%	186%	3.3%	72%
	2020	4,024	96,534	29,344	30%	28,797	98%	30%	1.6%	32%
	2021	1,581	114,394	21,642	19%	21,667	100%	19%	3.5%	21%

OBF. Table 1-3 tracks the process from claimed to evaluated savings for the OBF program. Evaluated realization rates are below 100% because there were issues in the calculation of savings (PY2020 gas) or in the application of exclusions, the process by which participants are assigned savings rather than receiving their performance-based savings estimates (both PY2021 results). Given the change in third-party M&V providers and M&V methods, the issues with exclusions in PY2021 are the greater concern. We discuss the specifics in the report. Because an updated OBF M&V plan was not available for consideration until the finalization of this report, the clarification of exclusions will have to take place during the review of that M&V plan. Also of concern are issues related to the extreme range of participant size in the OBF program. This extreme variability in participant consumption is part of the explanation of poor relative precisions that raise questions regarding the statistical validity of the results. It is also not consistent with the NMEC rulebook requirement regarding permissible project types and raises the question of whether population NMEC methods can address non-routine events in the presence of this level of variability.

Table 1-3. Summary of OBF electric and gas gross savings, PY2020 to PY2021

Program year	Electric sites	First year savings (kWh)							
		Claimed savings	PA M&V results	PA M&V realization rate	Evaluated savings	Evaluation realization rate	Gross realization rate	Percent savings	Relative precision
2020	322	2,436,946	2,436,946	100%	2,466,287	101%	101%	1.7%	238%
2021	392	10,206,131	13,774,580	135%	8,596,574	62%	84%	7.0%	86%
Program year	**Gas sites**	**First year savings (therm)**							
		Claimed savings	**PA M&V results**	**PA M&V realization rate**	**Evaluated savings**	**Evaluation realization rate**	**Gross realization rate**	**Percent savings**	**Relative precision**
2020	28	5,317	148,573	2794%	51,741	35%	973%	2.6%	210%
2021	28	123,282	538,942	437%	423,454	79%	343%	80.7%	66%

1.4.1.2 Program attribution and net savings

Table 1-4 provides the program attribution for residential PY2019 to PY2021 P4P installations and PY2020 and PY2021 non-residential OBF installations. The NTGR results range from a low of 45% for the HomeIntel program to a high of almost 90% for Comfortable Home Rebates (CHR).

Table 1-4. Program attribution and NTG values for population NMEC programs, PY2020 and PY2021

Program	Sample	Free ridership score	NTGR	Relative precision +/-
P4P - HomeIntel	492	55%	45%	5%
P4P - Comfortable Home Rebates	596	11%	89%	1%
P4P - Home Energy Rewards	1,748	27%	73%	2%
On-Bill Financing Alternative Pathway	96	44%	56%	9%

P4P. The P4P NTGR estimates appear to vary with the degree of focus on equipment measures rather than behavioral interventions. The CHR program's NTGR of 89% is consistent with the substantial investments motivated by the program. The timing was the biggest driver of attribution for this program, indicating that the program accelerated the timeline of purchasing and installing EE measures compared to non-program scenarios. The Home Energy Rewards (HER) program, with a combination of direct install measures and behaviorally oriented offerings, received an NTGR of 73%. Online marketplace measures were part of the HER offering with smart thermostats, which had low free ridership, as the most purchased measure in the sample.[21] Finally, HomeIntel, a virtual energy audit program, received an NTGR of 45%. HomeIntel is a purely behavioral program that does not incentivize any EE technology. Because of the lack of measure-based savings, we based the program's NTGR solely on the respondents' self-reported likelihood of taking the program-recommended action(s) in the absence of the program rather than timing, efficiency, and quantity (TEQ) questions.

OBF. The OBF program received an NTGR of 56%. This value is lower than NTGRs reported in previous evaluations but reflects a diverse and extensive sample with 99 measure-specific responses (representing HVAC, lighting, and refrigeration). Prior NTGR evaluations had limited or imbalanced samples. In one previous evaluation, a single project with 80% of the program savings in the survey sample received a 100% program attribution score - dramatically raising the program NTGR.

1.4.1.3 Total savings

The PA final claimed savings reflect true-ups based on metered M&V results filed with the CPUC and provided in CEDARS.[22] The PA revised its M&V results, to account primarily for the effect of the pandemic.[23] DNV's evaluated results are based on the replication and validation of the revised PA M&V savings. We compared the evaluated values to the final claimed savings to generate gross realization rates.

Table 1-5 provides the final electric claimed savings (final claimed gross savings), the evaluated gross savings, and the gross realization rates for each population NMEC program. The evaluated gross savings are adjusted to reflect the influence of the programs using net-to-gross ratios (NTGRs) and generate net savings. Our evaluation indicates that the population NMEC programs in combination caused 0.3, 2.5, and 6.7 million kWh of electric savings in PY2019, PY2020, and PY2021, respectively.

Table 1-5. Population NMEC electric claimed and evaluated savings, PY2019 – PY2021

Program	Program year	Final claimed gross savings (kWh)	Evaluated gross savings (kWh)	Gross realization rate	Evaluated NTGR	Evaluated net savings (kWh)
P4P - HomeIntel	2019	595,895	391,318	66%	45%	176,093
	2020	643,945	561,825	87%	45%	252,821
	2021	1,994,086	1,185,409	59%	45%	533,434
	2019	2,883	57,373	1990%	89%	51,062

[21] Discounts for smart thermostats through the online marketplace were as high as 60%. In addition, the program offered rebates for these devices, which allowed most participants to obtain them for free. This incentive level likely reduced the acquisition barrier sufficiently and led to high attribution for the measure.
22 PY2019 final claimed savings do not reflect true-ups because the PA did not file these.
23 The PA M&V savings for all the P4P programs diverge substantially from the final claimed values. The PA M&V P4P electric savings are up to seven times lower than the final claimed values.

Program	Program year	Final claimed gross savings (kWh)	Evaluated gross savings (kWh)	Gross realization rate	Evaluated NTGR	Evaluated net savings (kWh)
P4P - Comfortable Home Rebates	2020	638,753	860,447	135%	89%	765,798
	2021	3,201,543	1,383,443	43%	89%	1,231,264
P4P - Home Energy Rewards	2019	38,085	102,994	270%	73%	75,186
	2020	429,650	155,597	36%	73%	113,586
	2021	673,671	107,230	16%	73%	78,278
On-Bill Financing Alternative Pathway	2020	2,436,946	2,466,287	101%	56%	1,381,121
	2021	10,206,131	8,596,574	84%	56%	4,814,081

Similarly, we compared the evaluated to the final claimed gas savings to generate gross realization for population NMEC gas programs.[24] Table 1-6 provides the final gas claimed savings (final claimed gross savings), the gross realization rate, and the gross evaluated savings for each population NMEC program. Evaluated gross savings are also adjusted to reflect the portion of the savings that can be attributed to the program using net-to-gross ratios (NTGRs). Our evaluation indicates that the evaluated population NMEC programs, in combination, caused approximately 18, 89, and 341 thousand therms of gas savings in PY2019, PY2020, and PY2021, respectively.

Table 1-6. Population NMEC claimed and evaluated gas savings, PY2019 – PY2021

Program	Program year	Final claimed gross savings (therm)	Evaluated gross savings (therm)	Gross realization rate	Evaluated NTGR	Evaluated net savings (therm)
P4P - HomeIntel	2019	39,163	23,025	59%	45%	10,361
	2020	79,960	15,176	19%	45%	6,829
	2021	39,734	28,797	72%	45%	12,959
P4P - Comfortable Home Rebates	2019	305	994	326%	89%	885
	2020	24,052	36,358	151%	89%	32,359
	2021	54,751	83,818	153%	89%	74,598
P4P - Home Energy Rewards	2019	4,768	8,847	186%	73%	6,458
	2020	96,534	28,797	30%	73%	21,022
	2021	114,394	21,667	19%	73%	15,817
On-Bill Financing Alternative Pathway	2020	5,317	51,741	973%	56%	28,975
	2021	123,282	423,454	343%	56%	237,135

1.4.2 Application of NMEC requirements

In our evaluation of the application of NMEC requirements, we focused on how the programs conformed to rules set forward in five key areas in the NMEC Rulebook. These findings are primarily based on information available through early 2023, including interviews with program staff. The PA updated its P4P and OBF M&V plans in May 2023 and August 2023, respectively, which are not fully reflected in this evaluation.

Participant eligibility criteria. Participation eligibility requirements are crucial in NMEC programs as they ensure that participants have the data needed to support a performance-based savings estimate and do not engage in activities that could affect savings estimates via unaccounted-for non-routine events (NREs). All four programs require the customer to have at least 12 months of baseline period data. The P4P programs also flag and address the addition of electric vehicles (EVs), solar panels, battery storage, or participation in other programs.[25] Because the OBF program pre-dates the NMEC Rulebook and the updated M&V Plan became available after DNV completed its evaluation, it is not fully clear how OBF addresses the addition of solar or EV. It is clear, for example, from the code replication that, in PY2021, some sites with an

[24] Except for the Comfortable Home Rebates program, the PA M&V P4P gas savings are generally lower than (up to six times) the final claimed values.
[25] The addition of equipment, generation, or load-modifying activities will be conflated with savings in the pre-post comparison.

undefined solar flag receive assigned savings rather than a performance-based value. More problematically, the PA M&V process removed some sites without explanation, and as a result, savings magnitude and statistical accuracy improved substantially. Such ad hoc exclusions can pose a meaningful risk to population-level NMEC measurement. M&V plans must be up to date and include a complete description of all exclusions in the M&V code, the reasons for application, and implications regarding ineligibility (assigned savings, zero savings, etc.).

Permissible project types. The NMEC Rulebook requirement regarding population NMEC Permissible Project Types states that "sites can reasonably be expected to have similar types of equipment holdings, as well as drivers and levels of energy consumption." (NMEC Rulebook 2.0, II.2.B.2). While the P4P programs typically serve homogeneous residential customers, the non-residential customers in the OBF program do not fit this description. For example, PY2020 OBF participants have pre-participation consumption ranging from greater than 13,000 MWh to less than 1 MWh and with preliminary savings ranging from 4,667 MWh to 2.5 MWh.[26] The "similar types" of sites allow standard aggregate consumption data analysis techniques to produce acceptable savings precision estimates while addressing NREs. The OBF program's aggregate savings estimates are not statistically different from zero, let alone meeting the higher standard of relative precision of 90/25. This appears to be a direct result of unexpected negative site-level savings results (likely NREs) combined with the extreme variability in size and savings in the population. The OBF program has instituted size constraints over time. This evaluation's results indicate customer size is an important consideration, and those constraints in place at the time may not have been sufficiently stringent. There is a parallel but wider policy question regarding what levels of savings programs can claim using population NMEC versus site-NMEC methods, which afford a greater scrutiny of the site-level process.

Incentive payment structure. The incentive payments to all four evaluated programs appear to follow NMEC rules. The NMEC Rulebook requires that at least 50% of program payments to implementers reflect payable savings calculated using population-level NMEC methods and recommends that 100% of such payments reflect estimated savings using these methods. All residential population NMEC program payments to implementers are based entirely on metered savings, except for the CHR program, where 50% of the payments go to participant incentives. In the case of the OBF program, the PA providesa no-interest loan to a participating customer based on estimated monthly energy savings, and the customer pays the loan amount regardless of performance. Thus, considering whether incentive payments reflect performance is not necessary for this program.

Qualifying measures. The NMEC Rulebook is not prescriptive regarding qualifying measures and allows "measures currently allowable through the deemed and custom energy efficiency programs." (NMEC Rulebook 2.0, II.2.E.1). Allowed measures include behavioral, retro-commissioning, and operational. Furthermore, the PA Advice Letters encouraged programs to incorporate multiple measures and interventions that meet customer needs and can produce savings. For the residential programs, the PA described in the PIPs and Advice Letters the set of measures programs targeted, which are consistent with expectations for population NMEC programs. As discussed above, the design of the OBF program pushes the bounds of Rulebook definitions for population NMEC, and measures play a role in that.

M&V plans. As part of our evaluation of the Rulebook's M&V requirements, we used the descriptions of the M&V plans provided in the Advice Letters, the PIPs, the PA Early M&V report for PY2019 P4P programs,[27] the M&V code set, and information from PA program staff interviews. To the extent possible, we also reviewed the information provided in the latest OBF and P4P M&V plans, released in May and August 2023.[28] The programs fulfill the Rulebook requirement that population NMEC programs have M&V plans in place to estimate the energy efficiency impacts. However, the PA did not

[26] Three sites had preliminary savings expectations greater than their baseline consumption.

[27] Demand Side Analytics, "Early M&V Report for Program Year 2020 Residential Pay-for-Performance Program." Demand Side Analytics, July 12, 2022. https://pda.energydataweb.com/api/view/2653/Early%20M%26V%20Report%20PY%202020%20PGE%20Res%20P4P%2007-12-2022%20Draft%20Final.pdf.

[28] The 2023 updates are substantially different and bear little resemblance to the early plans.

satisfy the current Rulebook requirement that M&V plans get set before the start of such programs and applied consistently once the programs are underway. However, the Rulebook exempted pre-existing programs, including the P4P and OBF programs, from strict adherence to updated rules.

Program design criteria. The NMEC Rulebook specifies that "Population-level NMEC program designs must meet or exceed 90% confidence / 25% range Fractional Savings Uncertainty (FSU) [29] as calculated using the ASHRAE[30] methods at the daily level or using other methods that achieve at least the same levels of certainty." (NMEC Rulebook 2.0, II.2.C.1). The NMEC Rulebook is not sufficiently clear that this criterion refers to aggregate program level FSU in keeping with standard billing analysis precision measurements. The reference to ASHRAE methods, which are explicitly site-level calculations, could be misinterpreted as site-level eligibility rather than aggregate requirements. An aggregate FSU is more consistent with ensuring NREs are addressed appropriately in the population NMEC context. On the other hand, the program implementation plans did not provide how they predicted FSU.

1.4.3 Evaluability of Population NMEC programs

DNV identified three steps required for a successful evaluation of a process with an embedded M&V.

1. Replication of the embedded M&V process to confirm that it produces the claimed results.
2. Assessment of the methods used to ensure they:
    a. Meet basic NMEC requirements.
    b. Produce program-specific savings that are robust and reasonable.
3. Uniformity in claimed and PA M&V savings.

Replicability and transparency of PA M&V results. The essential first step of the NMEC evaluation is replicating calculated savings and an in-depth assessment of the process used to estimate the savings. The NMEC Rulebook states that "the methods used to calculate savings for NMEC programs must be documented in the program-level M&V sufficiently such that savings calculations can be replicated by the PAs as well as the Commission and its impact evaluators." (NMEC Rulebook 2.0, III.3.A.5). The Rulebook also says, "All analytical methods, including tools, algorithms, and software used in savings and incentive or compensation payment calculations, must be made available to Commission staff and its consultants upon request." (NMEC Rulebook 2.0, III.3.A.2).

Despite the guidance from the Rulebook, replicating the population NMEC results alone was overly time-consuming and, in some cases, not entirely possible. The current evaluation is the tail-end of a long process, which required three separate replication endeavors. During this process, DNV received incomplete and different analysis data than the PA appeared to have used, resulting in partial replication success. The analysis data also included duplicated energy consumption for a third of P4P participants, which required correction. In addition to correcting the data, the PA instituted a complete overhaul of the M&V approach. Thus, the final replication required recoding and substantially overhauled analysis and reporting. The PA also used different methods and software to generate PY2020 OBF and PY2021 OBF results and provided different levels of documentation of the approaches, adding to the challenge. The replication of the population NMEC results was so challenging that this aspect of the evaluation overshadowed the remaining evaluation efforts.

Assessment of the M&V approach used to determine savings. The program implementer makes decisions at every step of the performance measurement that will have implications for estimated savings. The replication of PA M&V findings makes it possible to validate each of those decisions made throughout the program performance assessment and determine

[29] Fractional savings uncertainty (FSU) is similar to relative precision in that it measures the uncertainty around the expected savings. As the value of FSU decreases, confidence in the estimated savings level increases.

[30] ASHRAE (American Society of Heating, Refrigerating, and Air-Conditioning Engineers) is a national standard-setting entity. It provides guidelines used to determine energy efficiency savings.

the implications for final savings estimates. This is the validation process where the performance estimate process is compared to best evaluation practices.

The models used for the P4P and OBF programs meet the basic methods guidance of the Rulebook and conform with industry standards. However, some aspects of the methodology used to evaluate all programs are novel in the energy context and deserve additional attention. Most analysis elements in the evaluation, such as data preparation and data sufficiency checks, also appear to meet accepted criteria. However, others, such as outlier treatment and customer attrition documentation, require additional attention. With the challenges related to replication for this evaluation, not all decisions were directly observable, making it difficult to assess all analysis elements appropriately.

Uniform claimed and PA M&V savings. The NMEC Rulebook defines claimable savings as "the savings reported by Program Administrators to the Commission before formal evaluation, measurement, and verification (EM&V)." (NMEC Rulebook 2.0, IV Definitions). For the NMEC program, PAs file initial savings claims for programs and adjust these based on metered-based savings via true-up claims. Initial claims are reported for each program year while the trued-up values, negative or positive, are reported in subsequent program years. DNV calculated final claimed savings for each program and program year, by summing initial claimed and trued-up savings reported in CEDARS.

However, the revised PA M&V results and the supporting data that DNV received did not match the PA's savings claims in CEDARS. As a result, DNV had to consider which savings numbers to use as the basis for evaluation – the CEDARS claims or the revised PA M&V. We elected to use the PA M&V values. These differing savings numbers complicated the process. A successful evaluation needs clear, accurate, and timely M&V savings that reflect those claimed with the CPUC.

1.4.4 Program effectiveness

This evaluation covers four distinct programs with unique program designs. The performance-based savings estimates and the evaluation survey provide insight into where these programs succeeded. The P4P programs, all targeting residential customers, motivated more generous savings compared to similar programs previously implemented that did not use NMEC.

Two of the P4P programs (HomeIntel and Comfortable Home Rebates) delivered notable energy savings, with average household electric savings of 5-8% and gas savings of 3-9%. The third P4P program (Home Energy Rewards), initially designed to include contractor-driven EE interventions, primarily provided small energy-saving measures due to the COVID pandemic. This program achieved more modest but expected electricity savings of 1-3% and gas savings of 2-4%.

- As a behavioral program, HomeIntel educated participants about their energy use, compared these to similar homes, and offered customized EE recommendations based on their energy use profiles. It targeted high-opportunity customers to achieve the noted savings. Survey results indicate HomeIntel participants were affluent with high educational attainment, which may have enabled them to navigate program information to achieve desired program outcomes. They also lived in older and larger homes, which could afford high savings opportunities.[31] While this program was successful in delivering savings, two sources of dissatisfaction noted by a subset of participants were: (1) that they were already implementing some of the program-recommended actions (2) and that some of these recommendations or installations were too expensive.
- Comfortable Home Rebates offered two participating pathways that either retrofitted homes or tuned up their HVAC systems. It targeted customers based on building characteristics, summer energy use, and inland climate zones where program installations were likely to succeed the most. Retrofits included HVAC system replacement, while HVAC tune-ups focused on optimizing the existing system. Savings as a percentage of consumption have increased each year to

[31] Behavioral programs such as Home Energy Reports generally motivate up to 3% of energy savings. Thus, even a 5% reduction is significant for such programs, but the 7% to 8% savings achieved by HomeIntel are substantial, particularly given the average size of participating homes. These achievements are outside of the norm for behavioral programs and deserve scrutiny.

notable levels for this program. Other programs offering HVAC optimization have not delivered the expected savings in recent years, and this program appears to have solved some of these challenges.

- The Home Energy Rewards provided online audits, energy-saving kits, and discounted energy management measures through an online marketplace that have helped the program achieve the expected modest energy savings. Like HomeIntel, the program targeted high-opportunity customers who were affluent and highly educated. However, its inability to provide contractor-driven retrofits because of the COVID pandemic may have limited its savings potential.

1.5 Conclusions and recommendations

Despite the relative newness of population NMEC programs, impact results provide evidence of the potential of the approach. All population NMEC programs faced challenges with at least some aspects of the embedded evaluation methodologies and translating those results into saving claims. However, this evaluation demonstrates that the performance-based programs delivered by the PA were consistent with most NMEC Rulebook expectations and the savings validated for the residential P4P programs were generally better than savings from similar programs implemented recently in California. Furthermore, while there were various challenges, the overarching goal of developing empirically-based savings estimates that minimize ratepayer risk was met. In total, these results represent a successful, if tentative, step to demonstrating the substantial potential of NMEC programs for California.

Clarify necessary steps to take population NMEC to the next level. This evaluation considers the first programs developed under population NMEC (or ported over from HOPPS) and looks at a period impacted by the COVID pandemic, a systemic NRE unlike any previously seen. Concurrent with this process, a Working Group provided feedback on the existing NMEC Rulebook v.2.0, and revisions are underway. In the context of these developments, change, and external stresses, it is possible to focus on basic steps that will move NMEC to the next level.

Recommendation:

- Require up-to-date program implementation plans, program M&V plans, and final M&V reports prior to evaluation.
- Require timely savings claims in CEDARS consistent with internal M&V results.
- Require a package of internal M&V code and data documented to make evaluator replication straightforward.
- Offer more explicit guidance on eligibility requirements, for example, no addition of solar generation during the program period.

Explore and address possible risks in the NMEC process to ensure reliable and robust NMEC programs going forward. Many of the issues identified in this evaluation can be explained by the basic technical challenge of embedding the M&V function as part of the program implementation and the unprecedented challenge to both program implementation and evaluation caused by COVID-19. There remain areas of potential risk that could be problematic for all parties involved that deserve further and ongoing attention.

Recommendation:

- Rules and the application of rules need to continually evolve to address challenges related to the precision of savings estimates and the potential for misuse of NMEC methods**.**
  - Programs that use population NMEC methods should demonstrate that they can appropriately address the full range of NRE risks (such as changes in building occupancy) that could lead to potential over- or under-estimation of savings.
  - Rules that address customer population variability should be in place to address the possibility of large customer NREs that undermine a program’s savings and precision. This may require redefining the FSU calculation for

population NMEC. More generally, this includes developing methods for identifying strategic NREs if they occur and addressing them appropriately.
  - There should also be rules that address new onsite solar during the baseline or performance period for OBF.

- The suitability of NMEC hourly savings for the application of avoided cost shapes for 2024 needs to be fully vetted.
  - While this evaluation focused on kWh and therm claims, in 2024, all electric claims will be based on hourly results. This shift requires a thorough examination of hourly savings methods and results to determine suitable approaches for the evaluation of total system benefits.
  - In addition, there ought to be a focus on the appropriate precision level for hourly, including peak demand savings estimates. The precision level should account for the limited number of hours over which peak savings estimates can be made and the portfolio size required to achieve these.

Program effectiveness. In general, the P4P programs appear to have delivered notable savings, particularly in light of recent evaluations that indicate lower savings achieved by similar non-P4P programs. Part of their success seems to be due to the more effective targeting of participants likely to maximize savings.

Recommendation:

- Despite the evident success, customer feedback indicates room for improvement in targeting messages to what is present at a participant's home and what the customer is willing to invest.

2 INTRODUCTION

This report presents DNV's PY2019 to PY2021 evaluation of projects based on population NMEC methods on behalf of the California Public Utilities Commission (CPUC). NMEC is a set of statistical tools and approaches that help determine the energy consumption impact of EE programs based on pre-and post-intervention meter data. Programs that estimate energy savings using NMEC (NMEC programs) claim final savings based on measured impacts at the meter rather than on individual measure deemed or calculated results.

2.1 Evaluation background

NMEC originated from two bills (AB 802 and SB 350) that authorized utilities to deliver pay-for-performance programs based on meter-based NMEC analysis.[32] There are two types of NMEC programs: site-specific and population-level. This report focuses on population NMEC programs where the same method can be applied uniformly to all participants.[33]

Population-level NMEC projects are a focus of our evaluation efforts because NMEC is a new and essential part of California's strategy to expand and deepen EE programs in the state. Given the overall population of NMEC projects is still relatively small, this evaluation encompasses much of the population NMEC efforts (including participation) from 2019 through 2021. This allows us to include all programs and their different strategies and target populations.

Because NMEC programs or projects have an embedded performance measurement process, the role of this evaluation is expected to be different than the evaluation for other energy efficiency programs. It needs to validate the embedded performance measurement and assess the overarching NMEC process (as defined in the Rulebook) to ensure it is consistent with the Rulebook and evaluation best practices. That is, early NMEC evaluations will serve as both evaluations of specific NMEC programs and, to the extent possible, the NMEC process embodied in those programs. Once the NMEC process is fully mature, the role of evaluation for an NMEC program/project may be simply to validate that the embedded process worked as intended.

2.2 Evaluation purpose and objectives

For this evaluation, we validated the gross and net savings of population NMEC programs with embedded performance measurements, assessed the processes, methods, and application of NMEC program requirements, and conducted program participation characterization. We determined the programs' application of NMEC requirements by referring to the NMEC Rulebook, which includes the CPUC's specific requirements for NMEC programs and measurement and verification (M&V) plans.

The purpose of the evaluation was to:

- Validate gross and net kWh and therm savings for population NMEC projects with claims in PY2019 through PY2021.
- Evaluate the application of NMEC rules and adherence to population NMEC program requirements.
- Determine the conditions for a successful evaluation of population NMEC programs.
- Characterize program participation.

[32] The 2015 California Assembly Bill (AB) 802 stated that the CPUC shall "authorize electrical corporations or gas corporations to provide financial incentives, rebates, technical assistance, and support to their customers to increase the energy efficiency of existing buildings based on all estimated energy savings and energy usage reductions, taking into consideration the overall reduction in normalized metered energy consumption as a measure of energy savings."

[33] Programs that apply population NMEC methods are evaluated under Group A since these closely resemble other evaluation methods that typically fall under this contract. Programs that use site-specific NMEC methods fall under Group D.

2.3 Evaluated programs

Table 2-1 lists the population NMEC programs we studied as part of our evaluation, including names, program IDs, and a description of each program.

Table 2-1. Population-based NMEC programs, PY2019 - PY2021

Program name and ID	Program description
P4P - HomeIntel Residential Program (PGE_Res_001b)	HomeIntel is a Pacific Gas & Electric (PG&E) residential downstream program implemented by Home Energy Analytics (HEA). The program audits the energy use profile of participating homes to educate participants about their energy use and offer them no or low-cost energy savings tips. The program's support staff also help participants to implement energy saving practices and solutions.
P4P - Comfortable Home Rebates (PGE_Res_001a)	Comfortable Home Rebates is a PG&E residential downstream and direct install program implemented by Franklin Energy. The program targets single-family homes and helps participants reduce their energy use and bills and improve the comfort of their homes. The program offers a retrofit of the home through an improvement pathway or upkeep of the HVAC system through a maintenance pathway.
P4P - Home Energy Rewards (PGE_Res_001c)	Home Energy Rewards is a PG&E residential downstream program implemented by Franklin Energy. It offers customers energy audits, energy-saving tips, a free energy-saving kit, online discounts on energy-efficient products, and rebates for smart thermostats.
On-Bill Financing Alternative Pathway (PGE210911)	On-Bill Financing Alternative Pathway is a PG&E downstream non-residential program implemented by various vendors and trade allies and offers customers utility financing for all energy efficient upgrades including those outside of utility rebate and incentive programs.

2.4 Reported gross and net electric and gas savings

PAs that claim savings based on population NMEC methods report initial savings in CEDARS[34] based on forecast or workpaper values. They true-up these initial claims based on metered results once 12-months of post-installation data become available. The sum of the initial and trued-up claims provides the final claimed savings for each program and year. The population NMEC programs we are evaluating followed this process for program years 2020 and 2021.[35]

We provide a summary of the reported initial claimed savings and any trued-up savings values associated with each program in these years in Table 2-2.

Table 2-2. Population-based NMEC programs savings claims, PY2019 to PY2021

Program and Program ID[36]	Program year	Electric sites	Total gross savings (kWh)			Gas sites	Total gross savings (therm)		
			Initial claimed	True-up	Final claimed		Initial claimed	True-up	Final claimed
P4P - HomeIntel (PGE_RES_001B)	2019	404	595,895		595,895	462	39,163		39,163
	2020	818	168,747	475,198	643,945	789	58,394	21,566	79,960
	2021	2,648	1,901,918	92,168	1,994,086	2,516	56,239	-16,505	39,734
P4P - Comfortable Home Rebates (PGE_RES_001A)	2019	562	2,883		2,883	43	305		305
	2020	2,053	218,433	420,320	638,753	1,346	8,272	15,780	24,052
	2021	2,276	946,456	2,255,087	3,201,543	2,002	34,252	20,499	54,751

34 California Energy Data And Reporting System.

35 The PA did not true-up the PY2019 P4P program claims and total claimed savings are based on the initial claimed savings for this program year.

36 The P4P program names and IDs in PY2019 were HEA Whole Home Savings (PGE210010), BIG Whole Home Savings (PGE210010), and Franklin Energy Whole Home Savings (PGE210010), respectively.

Program and Program ID[36]	Program year	Electric sites	Total gross savings (kWh)			Gas sites	Total gross savings (therm)		
			Initial claimed	True-up	Final claimed		Initial claimed	True-up	Final claimed
P4P - Home Energy Rewards (PGE_RES_001C)	2019	484	38,085		38,085	535	4,768		4,768
	2020	5,005	101,574	328,076	429,650	4,024	37,536	58,998	96,534
	2021	1,779	387,583	286,088	673,671	1,581	24,873	89,521	114,394
OBF-AP (PGE210911)	2020	322	27,012,587	-24,575,641	2,436,946	28	5,317		5,317
	2021	392	33,048,458	-22,842,327	10,206,131	28	123,282		123,282

2.5 PA M&V savings estimate updates

Since the application of NMEC methods to claim program savings is still in the early stages, the PA filed the trued-up savings to meet reporting deadlines but also undertook further analysis that accounts for the effect of COVID-19 on P4P program savings. The PA did not complete the study by the true-up filing deadline, and the updated M&V results were not reflected in CEDARs claims. DNV based its evaluation on these revised PA M&V results, rather than on CEDARs claims.[37]

Table 2-3 provides the final claimed savings reported in CEDARS and the updated PA M&V savings for the population NMEC programs in PY2019 to PY2021.

Table 2-3. Population-based NMEC programs savings claims and PA M&V results, PY2019 to PY2021

Program and Program ID	Program Year	Total gross savings (kWh)		Total gross savings (therm)	
		Final Claimed	PA M&V Results	Final Claimed	PA M&V Results
P4P - HomeIntel (PGE_RES_001B)	2019	595,895	391,318	39,163	23,056
	2020	643,945	561,852	79,960	14,479
	2021	1,994,086	1,185,439	39,734	40,669
P4P - Comfortable Home Rebates (PGE_RES_001A)	2019	2,883	57,373	305	981
	2020	638,753	861,410	24,052	36,531
	2021	3,201,543	1,386,773	54,751	83,720
P4P - Home Energy Rewards (PGE_RES_001C)	2019	38,085	102,994	4,768	9,021
	2020	429,650	155,945	96,534	29,344
	2021	673,671	108,948	114,394	21,642
On-Bill Financing Alternative Pathway (PGE210911)	2020	2,436,946	2,436,946	5,317	148,573
	2021	10,206,131	13,774,580	123,282	538,942

[37] The PA did not revise its PY2020 OBF with trued-up savings. Thus, DNV based its evaluation on the final claimed savings that reflect the true-ups reported in CEDARS for this program in PY2020.

3 METHODOLOGY

3.1 Data collection

In this section, we provide the data sources used in the evaluation, the primary data collection methods, and information on interviews we conducted with program implementers and the PA.

3.1.1 Data sources

We summarize the various data sources and the purpose of their inclusion in the evaluation in Table 3-1 below.

Table 3-1. Summary of population NMEC PY2019-PY2021 evaluation data sources

Data sources	Description	Purpose in evaluation
Program tracking data	PA claims filed with the CPUC that include program name and type, savings per claim, total claimed savings, etc.	To determine program information for evaluation
Detailed program participation data	Site-level information including customer names, addresses, and contact information, claim IDs, installation dates, details on installed equipment, and claimed savings	To identify participating sites (including customer and premise IDs), fuel type, claimed savings per site, installation dates, contact information for end-user surveys
PA M&V information	Methods, code, raw and intermediate data, site-level results from the PA M&V activity	To validate the energy savings claims of population NMEC programs
Energy use data	Customer hourly electric and daily gas interval advanced metering infrastructure (AMI) data	To estimate energy savings and validate claimed savings
Granular load profile data	Average hourly electric and daily gas AMI data for non-participating customers aggregated by solar status and climate zone (residential), and industry, climate zone, and size (non-residential)	Used to control for the effect of exogenous factors on energy consumption change post-intervention
Weather data	Hourly temperature for the relevant analysis period and locations, and typical meteorological year (normal) weather data	To weather normalize energy consumption
Telephone/web surveys	Surveys of participating utility customers and PA program staff.	To determine net-to-gross ratios (NTGRs) and net savings and understand site characteristics

The following list defines the data sources identified in the table above:

Program tracking data. This provides claims filed in CEDARS, including any true-ups for population NMEC programs. Data were at the aggregate program level for each population NMEC program in each program year.

Detailed program participation data. DNV requested site-level program data associated with the program claims in CEDARS to get details on participation, including installation site addresses, participant names and contact information, installation dates, the specific measures installed, and savings claimed. Data were at the installed equipment level for each participating site.

PA M&V information. For the population NMEC programs that DNV evaluated, the PA provided detailed site-level results and the meter-based methods, including code or detailed description of the approach, used to generate the results. DNV used the information that it received to validate the program's claimed savings.

Energy use data. DNV obtained AMI energy consumption data from the PA to estimate and validate the ex-post savings reported for the population NMEC programs. The energy use data were at the customer account level.

Granular load profile data. DNV also received the aggregate energy consumption data of selected non-participants the PA used to control for the effect of non-program-related changes on energy consumption. These are bin-level average loadshapes of randomly selected non-participants used as comparators.

Weather data. The National Oceanic and Atmospheric Administration (NOAA) and climate zone CZ2022 reference temperature files (CZ2022) were the sources of the weather data used in regression models accounting for weather sensitivity.[38] CZ2022 provides typical meteorological year (TMY) weather data for select California weather stations that are useful for long-term weather normalization. Data were at the hourly level for each station.

Survey data. The study collected primary information from participants through telephone and web surveys to determine program attribution, and customer characteristics and experiences. The collected customer information was for the responding sample.

3.1.2 Primary data collection

This section provides the primary research methods we used to evaluate population NMEC programs, including the data collection, sample design, survey approaches, mode, and disposition.

3.1.2.1 Data collection

Table 3-2 summarizes our primary data collection efforts, including key details such as respondent group, type of information collected, sample frame source, mode, and sample size used to evaluate population NMEC programs.

Table 3-2. Summary of primary data collection efforts for PY2019-PY2021 evaluation of population NMEC programs

Survey type	Respondent group	Data collected	Frame source	Mode	Stratification approach	Sample size
PA	Utility staff and implementers	Program information (installed measures, incentive levels), Customer participation, Application of NMEC rules	Program tracking data	In-Depth Interviews	N/A	Census[39]
OBF	Non-residential (commercial, industrial, institutions (e.g., schools))	Program influence – NTG, Participant characteristics, Program awareness and experience	Detailed program participation data	Phone	Measures installed	n = 99
P4P	Residential	Program influence – NTG, Demographic data, Occupancy, Program awareness and experience	Detailed program participation data	Web	Program	Census[40]

3.1.2.2 Sample design

The population or sampling frame was the list of installations from 2019 to 2021 for P4P and 2020 and 2021 for OBF provided by the PA. We used installations from these years for primary data collection to inform program attribution, participant characteristics, and experience.

[38] National Oceanic and Atmospheric Administration Hourly Weather Data; California Energy Commission. “Building Energy Efficiency Standards.” energy.ca.gov, Accessed 5/16/2023. https://www.energy.ca.gov/title24/; CALMAC. “CALMAC California Weather Files.” calmac.org, Accessed 5/16/2023.http://www.calmac.org/weather.asp.

[39] We conducted interviews with three implementers who ran P4P programs and the PA program staff.

[40] We attempted a census for the research among residential participants with the aim of a minimum of 62 to 68 completes by program to achieve a relative precision of ±10% at a 90% confidence level.

The program information summarized in Table 3-3 provides the number of participants by survey type and program used in the evaluation. The number of participants in the table provides population sizes that are the sources of all the survey sample frames.

Table 3-3. Population NMEC PY2019-PY2021 survey types and number of participants

Survey type	Program name	Number of participants
Residential P4P survey	P4P - HomeIntel (PGE_RES_001B)	4,494
	P4P - Comfortable Home Rebates (PGE_RES_001A)	3,024
	P4P - Home Energy Rewards (PGE_RES_001C)	7,513
Non-residential OBF survey	On-Bill Financing Alternative Pathway (PGE210911)	833

For primary data collection, our sample design approach was based on a census for P4P and a stratified sampling approach for OBF. Under the census approach for P4P, we attempted to collect data for all individual participants in the population with contact information. We aimed to get a minimum of 62 to 68 completes by program to achieve relative precisions of ±10 at 90% confidence levels. Under the stratified sampling approach for OBF, we selected sample units for the study from installations stratified by savings measured in MMBtu, which is the sum of kWh and therm savings converted to MMBtu. We then estimated the appropriate sample size for the program to achieve a targeted relative precision (±10%) at the desired level of confidence (90%). Once the sample size was calculated, we randomly chose primary sample points from the population based on the stratification plan. In addition, we selected a backup sample in case any sample points needed to be replaced. Replacement happened with sites that could not be contacted or evaluated. Appendix E provides the details for both P4P and OBF.

The primary objective of the surveys was to develop estimates of free-ridership. The survey data also provided information on participant experience and characteristics, including participant demographics and firmographics. The survey instruments and interview guides used for data collection are provided in Appendix F.

Table 3-4 summarizes topics covered by the residential P4P and non-residential OBF surveys.

Table 3-4. Population NMEC survey topics among program participants, PY2019 to PY2021

Survey topics	Residential P4P	Non-residential OBF
Free-ridership		
Equipment verification	●	●
Free-ridership questions	●	●
Program Outreach and Participation		
Drivers of program participation	●	●
Reasons for program loan		●
Effectiveness of rebates and incentives	●	●
Program information source		●
Program Experience/Satisfaction		
Information Provided	●	●
Perceived program benefits	●	●
Program satisfaction	●	●
Barriers	●	●
Clean Tech Adoption		
Current use of clean energy products	●	●
Interest in clean energy products	●	●

Survey topics	Residential P4P	Non-residential OBF
Demographics/Firmographic		
Home/Facility ownership status	●	●
Home/Facility size	●	●
Home/Facility age	●	
Home/Facility type		●
Number of employees		●
Participant characteristics - language, income, primary language	●	

3.1.2.3 Survey mode and disposition

P4P surveys. We conducted a web survey among participants who received measures from the three P4P programs, HomeIntel, Comfortable Home Rebates, and Home Energy Rewards, which targeted mostly single residential customers.

We conducted a web survey among program participants over approximately 2 weeks from April 23 to May 5, 2023. The sample frame for this survey was the list of participants for whom email addresses were available. We offered five respondents a chance to win a $300 gift card to complete the survey and sent four reminders to encourage invitees to complete the survey.

Table 3-5 summarizes the sample disposition for the residential survey. DNV adopted proven best practices in fielding this survey, including:

- Providing respondents with a link to validate the legitimacy of the survey effort.
- Co-branding the survey with the CPUC and IOU logos.
- Issuing the survey invitation from an email address with the IOU domain to reduce the likelihood of it being caught by the respondent's spam filters.
- Providing a letter from the CPUC study manager on the importance of this research and customer responses to energy efficiency programs.

Table 3-5. Sample disposition for PY2019-2021 residential P4P web surveys

Program	Invites sent	Partially completed	Completed	Response rate
P4P - HomeIntel (PGE_RES_001B)	4494	180	954	21%
P4P - Comfortable Home Rebates (PGE_RES_001A)	3024	74	267	9%
P4P - Home Energy Rewards (PGE_RES_001C)	7513	246	1473	20%

OBF surveys. We administered non-residential surveys for the OBF program where site managers were the primary decision-makers for installations. We used phone surveys to collect data among this segment of participants for approximately two weeks starting on April 18, 2023, through May 2, 2023. DNV fielded up to six calls per site during this period. The objective of these surveys was to estimate free-ridership and develop the firmographic characteristics of the participants.

The sample frame for this survey was based on a list of PY2020 and PY2021 OBF participants selected to represent measure groups and savings levels provided by the program. We offered five participants the chance to win a $100 gift card as an incentive to complete the survey.

The sample disposition for the non-residential OBF survey is summarized in Table 3-6.

Table 3-6. Sample disposition for non-residential PY2020-PY2021 OBF phone surveys

Program	Attempted calls	Partially completed	Completed	Response rate
On-Bill Financing Alternative Pathway (PGE210911)	276	10	99	36%

3.1.1 Interviews with implementers and the PA

We conducted in-depth interviews with program administrators and implementers to understand the application of NMEC rules, program incentives, execution, and installed measures. The information we collected included the types of M&V methods in place, program eligibility criteria, incentives, data tracking, measure selection, rebates, and information offered to participants. Appendix F provides the PA and implementer guides we used.

We interviewed P4P implementers in March 2023 and PA P4P and OBF staff in April 2023. Table 3-7 provides the interview log, which includes the program names and call dates for each of the interviews.

Table 3-7. Population NMEC PY2019-PY2021 PA and implementer interview log

Population NMEC Program	Program Administrator call date	Implementer call date
P4P - HomeIntel (PGE_RES_001B)	24-Apr-23	27-Mar-23
P4P - Comfortable Home Rebates (PGE_RES_001A)	24-Apr-23	27-Mar-23
P4P - Home Energy Rewards (PGE_RES_001C)	24-Apr-23	24-Mar-23
On-Bill Financing Alternative Pathway (PGE210911)	24-Apr-23	NA

3.2 Impact

3.2.1 Gross savings approach

P4P. The PA population NMEC programs claimed savings for residential P4P programs with interventions in PY2019 through PY2021. All the programs claimed initial savings with the CPUC reported in CEDARS, while only the PY2020 and PY2021 programs provided trued-up savings. Final claimed savings for all the programs are the sum of the initial and trued-up savings reported in CEDARS. The PA claimed trued-up savings to meet reporting deadlines, then undertook an in-depth M&V process to account for exogenous changes, including the pandemic.

In this section, we provide our approach to replicate and validate the PA savings based on the results of this M&V process. We use the validated savings to generate evaluation realization rates (evaluated to PA M&V savings). These values indicate the success of our replication and validation of the PA M&V savings. The validated results are the basis of our evaluated savings. Additionally, we compare the PA M&V and evaluated savings to the final claimed savings (to obtain PA M&V and gross realization rates, respectively), which indicate the extent to which the programs achieved the final claimed savings in CEDARS.

As part of our replication of methods and results, we ran input data provided by the PA through translated code and, where possible, we compared intermediate outputs to corresponding provided data files. We also compared the final results after running the full process to results provided in M&V summary tables and reports to determine the replicability of the results, a central tenet of the NMEC approach.

For the residential P4P programs, we translated the Stata code used by the PA's vendor as part of our replication of the methods used to generate gross savings estimates. The Stata code we received included files that clean AMI and participant characteristics data; segment participant data; create electric and gas analysis datasets that combine AMI data, participant characteristics, granular load profiles, and weather data; run regressions to estimate hourly electric and daily gas site-level savings; define exclusions of site-level results; and roll-up the site-level results post-exclusion by program and program year. What follows is a simplified description of the gross savings approach. Details of the approach are provided in the program M&V plan.[41]

The M&V analysis provided in these files comprised several broad steps. The first step involved preparing electric and gas analysis datasets. To prepare these data, the code removes duplicate energy consumption reads and defines ineligibility criteria based on solar and other distributed energy resource (DER) status, participation in more than one program, and inadequate pre-participation data. The code also segments participants based on climate region (coastal, inland, north- and south-central valleys), solar status, and annual energy consumption level. The code combines the prepared AMI and participant characteristics data (which include climate region, solar and other DER status, and enrollment date) with granular profiles based on climate region, solar status, and energy consumption level.[42] The code then adds weather data based on the mappings of each participant to the nearest weather station. Finally, the code prepares regression datasets by defining pre- and post-installation periods, temperature splines, and a series of time indicators such as the day of the week and weekday used for modeling.

In the second broad step, the code runs site-level models that combine synthetic control and time of week temperature (TOWT) modeling approaches.[43] The models employ the granular profiles discussed above as the synthetic controls and the piecewise-continuous weather trend of the TOWT model to the address remaining weather correlation. The PA had originally estimated unique models for each hour of the week using this specification. The final PA M&V analysis, applied retrospectively to all prior years, estimates three seasonal synthetic control models for each customer and year. These models include dummy variables for all hours of the week, the appointed granular profiles, and a single piecewise-continuous weather trend across all hours. The definition of the seasons used to model energy consumption includes summer (June to September), winter (December to March), and shoulder (April, May, October, and November) periods. The code uses baseline period model coefficients and post-period variable values to generate predicted (counterfactual) energy consumption levels that reflect post-period conditions. These are compared to actual post-period energy consumption to estimate the impact for each hour or day by site. These "actual weather" estimated impacts, referred to as avoided energy use, are then separately regressed on weather data to weather normalize impact using CZ2022 typical temperature values.

In the third broad step, the code rolls up site-level hourly and daily weather-normalized values to the cohort level (all customers in an installation month) and by year for each program. An essential part of the roll-up is the exclusion of ineligible customer data (due to the addition of electric vehicles, solar PV, storage, tenant turnover, etc.) flagged in the data preparation stage. These customers' savings are replaced by average savings generated from the remaining eligible customers within groups (such as program year, program, and solar status) or set to zero, depending on the exclusion criteria discussed in detail in the results section. Results are then aggregated to the cohort level and program year for each P4P program.

[41] Pacific Gas and Electric Company Residential Pay-for-Performance Program Measurement & Verification Plan Version 2.1, PG&E. Revised: 05/31/2023

[42] The implementer constructed granular profiles as averages of 8760 hourly consumption values for segments from randomly selected non-participant customers. The segments reflected climate zone groups, solar status, and premise and load shape characteristics. Climate zone groups included coastal, inland, North Central Valley, and South Central Valley climate zones. Premise characteristic segments comprised customers with and without electric heat (identified using electric rate codes) and customers in consumption quartiles within each climate zone group. Load shape characteristics divided customers based on the percent of daily consumption between 1 pm and 4 pm to identify four occupancy patterns within each climate zone group, consumption bin, and solar status.

[43] The source of the original TOWT modeling approach is Mathieu, Johanna L., Phillip Price, Sila Kiliccote, Mary Ann Piette. 2011. "Quantifying changes in building electricity use, with application to Demand Response." IEEE Transactions on Smart Grid 2:507- 518. The original model featured rolling 3-month models that predicted load for the central month. It also featured an occupancy variable, derived from the data, allowing two distinct temperature trends. The term time of week and temperature model now appears to indicate any model utilizing a piecewise-continuous weather trend and hour of day and other time indicators.

OBF. This evaluation covers non-residential OBF program claims with project installations in PY2020 and PY2021. The programs claimed initial and trued-up savings reported in CEDARS. As for P4P, the PA undertook additional M&V analysis to estimate the savings of these programs. The M&V electric savings for PY2020 matched the final savings claimed in CEDARS, while PY2020 M&V gas and PY2021 M&V electric and gas savings differed from the final claimed savings. Again, like the P4P evaluation, we evaluated the M&V savings, which we compared to the final claimed savings in CEDARS to generate gross realization rates.

The PA used different third-party M&V providers, and different M&V methods, to estimate OBF program savings in PY2020 and PY2021. DNV replicated and used the different methodologies to validate claimed savings for each year. To validate the PY2020 results, DNV followed M&V documentation provided by the PA that outlined the approach, code snippets for data preparation, and a link to the publicly available modeling codebase, OpenEEmeter, used to model energy consumption. DNV also received the data needed to calculate PY2020 savings, including hourly electricity and daily gas consumption data for participants and granular load profiles. DNV used these sources to develop a set of codes to estimate OBF savings for each participant. Again, the following is a simplified description of the methods. Further information is in the OBF program M&V plan.[44]

The first broad step involved developing the analysis dataset using both participant data and granular load profiles. This step required checking and cleaning data based on the data quality requirements outlined in CalTRACK, including identifying any data gaps.[45] It also involved mapping each participant to a granular profile based on climate zone grouping (coastal or inland), industry group, solar status, and size or energy consumption level. NAICS codes facilitated the identification of industry groupings or similar businesses, while rate codes served to identify similarly sized sites. Such mappings allowed the selection of granular profiles composed of non-participants subject to comparable operating and weather conditions and whose energy consumption response is likely similar to the matched participants. The process mapped each granular profile to as many participants as was necessary.

The second step produced separate models for each participant and its matched granular profile based on pre-installation data and daily CalTRACK methods. These models estimated site-level energy consumption as a function of weather using a range of reference temperatures to identify the heating and cooling base points for each site and generated parameter estimates that indicate the levels of energy consumption correlated with heating degree days (HDD) or cooling degree days (CDD). They also identified the level of baseload not associated with either heating or cooling needs. DNV used the referenced, publicly available Python code, OpenEEmeter, to generate the site-level model estimates based on 365 days of pre-intervention energy consumption data (baseline data).

The site-level models developed using pre-intervention data were the basis of post-intervention counterfactual energy consumption estimates that reflect actual post-period weather for all participants and their chosen granular profiles. As for the P4P program, these produced “actual weather” avoided energy use estimates. These counterfactual values are compared to actual post-period energy consumption to estimate impact for each hour or day by site. The ratio between the granular profile model results, observed at post-installation weather, and the post-intervention granular profile then provided an adjustment for the participant model results, observed at post-installation weather.

The third broad step for 2020 OBF also paralleled the P4P process, with excluded sites being assigned savings based on the remaining, non-excluded sites. OBF exclusions included sites without the required 90% of pre-installation data or for whom baseline models could not be estimated, those with poor baseline models (CVRMSE > 1), and those whose percent unadjusted savings were outliers (above or below 50% based on the first 12 months of post-intervention data). The approach assigned savings to these sites by multiplying their ex-ante savings (engineering-based estimates of savings for

[44] On-Bill Financing Alternative Pathway Program-level M&V Plan. PG&E. Revised: 08/08/2023. https://cedars.sound-data.com/documents/download/2882/main/

[45] CalTRACK. “CalTRACK Methods.” docs.caltrack.org, Accessed 5/5/23. http://docs.caltrack.org/en/latest/methods.html#section-2-data-management

each site developed before implementation) by a 0.95 NTG ratio. However, the PA did not include the assigned savings in the true-up value reported in CEDARS. Because the PA only claimed savings for those sites without exclusions for PY2020, it was less important to pursue the specifics of those exclusions

For PY2021, the same third-party M&V contractor conducted M&V for both P4P and OBF. As a result, the PY2021 OBF M&V process was identical to that used for the P4P programs except for the ineligibility criteria and the assignment of savings to ineligible sites. For OBF, the process flagged sites as ineligible for electric and gas analysis if they had insufficient data. Additionally, it flagged sites as ineligible for electric analysis if their baseline models had poor statistical fit (CVRMSE > 1), their solar status changed, or the site accounted for more than 4% of the total savings across all projects. It then calculated realization rates for the eligible sites by implementer and program type (Lighting, Non-lighting, Both). To assign savings to the ineligible sites, the process involved multiplying their ex-ante savings by the implementer realization rate, if available, or the program type realization rate. Finally, results were aggregated across all eligible and ineligible sites to the cohort and implementer levels.

Parallel to P4P, we validated the PA savings based on the results of this M&V process and used the validated savings to generate evaluation realization rates (relative to PA M&V savings). These values indicate the success of our replication and validation of the PA M&V savings. Additionally, we compared the PA M&V and the evaluated savings to the final claimed savings (to obtain PA M&V and gross realization rates, respectively), which indicate the extent to which the programs achieved the final claimed savings in CEDARS.

3.2.2 Program attribution approach

This study also estimated what percentage of the gross energy savings claimed by the population NMEC programs are attributable to those programs (net savings). This analysis looked for evidence of free-ridership – claimed energy savings that would have occurred in the absence of the program. Gross measure savings estimate the change in energy use due to program participation, regardless of why customers participated, while net measure savings estimate the change in energy use without free riders.

We developed estimates of the ratio of net savings to gross savings, or NTGR, and then applied that ratio to gross savings to calculate net savings. An NTGR equal to 1.0 indicates that the programs influenced 100% of the energy savings. The difference between the measured NTGR and an NTGR of 1.0 is the free-ridership proportion. For example, 25% free-ridership yields an NTGR of 0.75, meaning 75% of the savings from program-incentivized actions or installations would not have occurred in the absence of the program.

For each of the three programs, we surveyed all participants about their likelihood of taking energy-saving actions or installing the energy saving measures on their own. We designated those who indicated that they were “very unlikely” to have completed the task or upgrade on their own as zero free-riders.

The timing question asks how soon each measure would have been installed absent the program. The program gets full attribution for any measure that would not have been installed at all, and it gets partial credit for accelerating the timing compared to when respondents say they would have installed the measure without the program.

The efficiency question applies to the efficient measures installed by the programs for which there is a standard efficiency version in the market. The efficiency question gives the program full credit for the measure if the respondents indicate they would have installed nothing or a standard efficiency measure instead of the efficient program measure. However, if respondents indicated they would have installed a high-efficiency measure, then the program received no credit for the measure. A third option, a mid-efficiency installation, resulted in partial program credit.

The quantity question asks how many units would have been installed absent the program. This question applies to measures where programs permit more than one installation per participating site. The quantity question gives the program credit if the respondents indicate they would have installed fewer measures absent the program.

In the sections below, we describe the methods we used to estimate levels of free-ridership for the P4P and OBF programs.

3.2.2.1 Program attribution for P4P programs

HomeIntel, a behavioral program, provided each participant an online audit and recommendations based on the audit. Participants reported a range of recommendations they received from the program following the audit. For this program, we asked participants how likely they would have been to complete energy-saving recommendations by the program on their own. We calculated the level of free-ridership and its complement -- the proportion of energy savings attributable to the HomeIntel program -- based on these survey responses.

For the CHR and HER programs, we asked participants about the likelihood of their taking energy-saving actions or installing energy-saving measures promoted by the programs on their own. If these participants said they were “very unlikely” to have completed the energy saving actions or measures on their own, we classified them as zero free riders. However, if they indicated there was some likelihood that that they might have completed these actions or measures on their own, we asked them additional questions about timing, quantity, and efficiency to determine free-ridership. Taken together, these dimensions allow for estimates of net energy (kWh, kW, and therm) savings attributable to each measure because those savings depend on the number of measures installed (quantity), the efficiency of the measures (efficiency), and when the measures are installed (timing).

Appendix D details how we scored participant survey responses to derive free-ridership values. We calculated measure-level NTGRs based on these, which we used to calculate measure- and program-level net savings.

3.2.2.2 Program attribution for OBF programs

OBF Program Attribution Principles and Approaches. OBF-AP programs allow utility customers to finance non-PA EE projects using only OBF. Before OBF-AP, OBF + Rebate projects used the CA Non-residential NTG protocol to determine program attribution.[46] This protocol considered both program influence and counterfactual scenarios to determine program attribution. Since the start of PG&E’s OBF-AP program in 2018, two studies have refined the approach to determine program influence and NTGRs for EE projects that primarily use program financing. A Cadmus study developed a program attribution approach for PG&E’s 2018-2019 OBF-AP projects using counterfactual scenarios after reviewing three established and salient NTG approaches.[47] An ODC evaluation of PY2018-2019 OBF programs also provided a program attribution approach based on counterfactual scenarios that take into consideration the likelihood that participants would have sought and secured alternative financing in the absence of the program.[48]

The Cadmus approach determined program attribution based on what participants said they would have done in the absence of OBF. It based program attribution scores on the probability that participants would have undertaken the project on their own, as well as on the timing, quantity, and efficiency (TEQ) of the project installations absent OBF. The approach estimated NTGRs by multiplying the probability of the participant undertaking a project/measure on their own (using a 0/1 binary factor with 0 assigned to participants who said they were unlikely to have done the project/measure on their own) by the TEQ scores.

---

46 California Non-residential Net-to-Gross Working Group. February 20, 2015. Methodological Framework for Using the Self-Report Approach to Estimating Net-to-Gross Ratios for Non-residential Customers. Prepared for California Public Utilities Commission, Energy Division

47 CADMUS, Evaluation of the On-Bill Financing - Alternative Pathway, PY2018-2019 FINAL REPORT. August 3, 2020. Cadmus. https://www.calmac.org/publications/OBF-AP_PY18-19_Process_Evaluation_Final.pdf

48 The ODC study covered 2018 OBF + Rebate and 2018 and 2019 OBF-AP projects, the latter of which excluded projects associated with survey respondents in the Cadmus study. PY2018–2019 California Statewide On-Bill Financing Impact Evaluation – DRAFT CPUC Group B. March 7, 2023. Opinion Dynamics. https://pda.energydataweb.com/api/view/2780/CPUC%20OBF%20Evaluation%20Report_DRAFT_03092023_clean.pdf

The ODC approach first considered whether participants sought or would have sought alternative financing and the probability of securing that financing. If the participants did not seek, or were unlikely to secure, financing other than OBF, the study determined program attribution using general financing counterfactual scenarios since there was no need to distinguish between the impacts of general financing and OBF in such cases. On the other hand, if participants sought, and were likely to get alternative financing, the study determined program attribution using OBF counterfactual scenarios. In either case, the NTGR reflected the likelihood of undertaking the project, and the TEQ of the installations absent financing. Unlike the Cadmus approach, the ODC study estimated NTGRs based on the average of the likelihood of undertaking the project without the program and the TEQ scores.

To determine program attribution, DNV considered the likelihood that participants would have used general financing (loans) to complete their projects in the absence of OBF. In cases where participants were not likely to seek general financing, there was no need to distinguish between OBF and general financing to determine program attribution. In such cases, DNV based program attribution on scores using general financing counterfactual scenarios. In cases where participants were likely to seek and/or secure general financing, we based program attribution on the participants' likelihood of undertaking the project in the absence of OBF.

Project attribution details. OBF tracking data was available at the project level. If possible, DNV used the mention of specific measures in the project description to ask respondents measure-level NTGR questions. We asked timing and efficiency NTG questions for all measures. We asked quantity NTG questions only for lighting and food service measures. The surveys asked participants the timing, efficiency, and quantity (if relevant) questions for up to three measures.

To ensure more robust results, we only calculated NTGRs for measures with five or more survey responses. After we dropped measures with fewer than five responses three measures remained: HVAC, Lighting, and Refrigeration. We then weighted the NTGR responses for each of these three measures back up to their respective populations. Table 3-8 summarizes, for each measure, the number of survey responses, the size of the participant population, whether we calculated attribution for the measure, and the weight (if we calculated an NTGR) for each measure.

Table 3-8. Population and respondent summary for OBF program participants, PY2020 and PY2021

Measure	Survey respondents	Participant population	NTGR analyzed	Weight
Boilers	2	4	No	
Controls	1	36	No	
Food Service	0	8	No	
HVAC	10	43	Yes	4.3
Lighting	80	754	Yes	9.4
Motors	2	22	No	
Other	2	16	No	
Refrigeration	6	37	Yes	6.2

The resulting NTGR calculations appear in section 4.1.1.2.

3.3 Application of NMEC rules

This section provides a summary of how DNV evaluated the application of the NMEC requirements by the population NMEC programs, most of which were launched before the NMEC Rulebook. We considered the following factors for this evaluation:

- Program elements, including targeted customer segments and eligibility requirements for participation, measures and program intervention approaches, and incentive payment structures

- M&V plans, including the proposed M&V approaches, barriers to implementing the approaches, program attribution plans, and cost-effectiveness plans
- Program management, including program plans to track information and program results, and available information to replicate trued-up savings

The application of NMEC for energy efficiency evaluation is still in the nascent stages. CPUC developed the NMEC Rulebook, which stipulates the requirements of the M&V process, program design criteria, permissible project types, incentives, and qualifying measures. Our evaluation examined how programs adhered to the different requirements laid out in the Rulebook. However, the plans and processes for the programs, which are provided in Advice Letters that the programs filed with the CPUC, precede the Rulebook. Our evaluation took this into account.

As part of the evaluation, we also examined the programs' Advice Letters (PG&E Advice Letter 3698-G/4813-E and PG&E Advice Letter 3697G/4812-E, issued March 25, 2016) and implementation plans (PIPs). Additionally, we conducted interviews with the PA (PG&E) program staff to gain insight into certain program aspects that were unavailable in the PIPs. Below is a list of all data sources/documentation we reviewed as part of our assessment of the application of NMEC rules:

- Rulebook for Programs and Projects Based on Normalized Metered Energy Consumption, Version 2.0, January 7, 2020
- PG&E Advice Letter for the Submission of High Opportunity Projects and Programs (HOPPs) Proposal - Residential Pay-for-Performance Program (3698-G/4813-E issued March 25, 2016)
- PG&E Advice Letter for the Submission of High Opportunity Projects and Programs (HOPPs) - On Bill Financing Alternative Pathway Program (3697-G /4812-E issued March 25, 2016)
- Program Implementation Plan, Comfortable Home Rebates, Franklin Energy, October 2021
- Program Implementation Plan, HomeIntel Residential P4P Program, Home Energy Analytics, May 2021
- Home Energy Rewards, Implementation Plan, Enertouch, Inc. dba Franklin Energy Demand Response, Version 1.1, October 2021
- On-Bill Financing – Alternative Pathway Implementation Plan, PG&E, May 19, 2021
- Early M&V Report for Program Year 2020 Residential Pay-for-Performance Program, Draft Final Report, Demand Side Analytics, July 12, 2022
- Early M&V Report for Program Year 2020 On Bill Financing-Alternative Pathway Program (OBF-AP), Population NMEC Sub-Program, Draft Final Report, Demand Side Analytics, July 12, 2022
- Evaluation of the On-Bill Financing – Alternative Pathway, PY2018-2019, Final Report, Cadmus, August 3, 2020
- Program Implementers interviews conducted on March 27, 2023
- PA Interviews conducted on April 24, 2023

Based on these data sources, DNV evaluated the program elements and methods used by NMEC programs to track and generate claimable savings. We assessed each program's M&V design and methodology in terms of its transparency, replicability, and soundness. These methods must meet industry best practices, must produce robust results, and must be applied to track ongoing program progress. We also noted any barriers to applying NMEC population methods for each program. We provide our findings regarding the application of NMEC rules in section 4.2.

3.4 Assessment of gross methods and results

We also examined the M&V process and results used to true-up population NMEC program savings. To address these, we examined if:

- If the M&V approaches are replicable and transparent

- If the methods are sound and meet industry best practices
- If elements of the analysis meet accepted criteria
- If the foundations on which results are based are within the scope of the NMEC Rulebook requirements
- If the true-up savings are reasonable

We provide our findings in section 4.3.

3.5 Characterize participation

We surveyed the population NMEC program participants based on participant rosters provided by the PA and survey instruments we developed to characterize participation in P4P and OBF population NMEC programs. We asked participants questions to gauge their program awareness, motivation for participation, satisfaction, demographics, and firmographics. Using the collected data, we determined installation rates of program measures among participants of two P4P programs, program participant characteristics, and customer experience in all PY2019 to PY2020 P4P and PY2020 and PY2021 OBF population NMEC programs. We provide our characterization of participation in section 4.4.

4 RESULTS

This section provides four areas of results. The impact section, section 4.1, provides gross impacts and net-to-gross findings. We combined these to provide total program savings. Next, in section 4.2, we assess the application of NMEC rules in the implementation and M&V of the population NMEC programs. We follow this with an assessment of the population NMEC programs' methods and results in section 4.3. Finally, in section 4.4, we discuss what we learned about the program participants from the surveys we conducted.

4.1 Impact

In this section, we present our findings on the impact of P4P and OBF population NMEC programs.

4.1.1 Gross savings

We provide gross savings results we used to evaluate the P4P and OBF program claims in this section.

4.1.1.1 P4P gross savings validation

Replication process. DNV replicated the P4P M&V results using the data and information provided by the PA. As detailed in the methods section, we used the information we received to generate regression datasets, synthetic control regression results, weather-normalized site-level savings, and final cohort-level savings. We received the following data and information for the replication:

- Raw customer characteristics data that included climate zone, solar and other distributed energy resource (DER) status, program enrollment information, and ineligibility criteria
- Raw interval data with hourly electricity and daily gas consumption
- Granular profile data with hourly non-participant bin-level energy consumption values by solar status and climate zone
- Weather data with actual and TMY hourly temperature data and mapping to participant sites
- Regression datasets that the PA used to generate site-level savings
- Site-level baseline energy consumption and savings at the hourly level for electricity and daily level for gas
- Site-level weather normalized baseline energy consumption and savings at the hourly level of electricity and daily level for gas

Figure 4-1 summarizes the process we used to replicate final cohort-level savings. We processed the customer characteristics data to flag observations ineligible for inclusion in the analysis. We joined the processed customer characteristics and interval usage data and cleaned this data (for example, by removing duplicate values) to produce the prepared data. We created additional variables, such as time indicators and temperature bins, and merged the data with the granular profile data to generate the regression data. We ran synthetic control and weather normalization regressions to generate site-level normalized savings.[49] We removed ineligible data from the site-level results and applied the average values from the remaining sites to fill savings for the excluded observations. Finally, we aggregated these savings to the cohort and program levels to produce the final savings results.

Figure 4-1. Steps used in PY2019-PY2021 P4P Results

Replication results. As indicated above, we processed the raw data to generate the regression data used for analysis. Table 4-1 presents the similarity of values of the fields from the DNV-prepared and the PA-provided regression datasets critical for the synthetic control regression and weather normalization stages. Approximately 99.7% of the energy

[49] The distribution of CVRMSE values from all site-level regressions is provided in Appendix G.

consumption and weather data in the DNV and PA regression datasets were the same. In addition, the granular profile consumption data used in the synthetic control regressions between the two regression datasets were equally similar.

Table 4-1. Comparison of DNV-prepared and PA-provided PY2019-PY2021 regression analysis datasets

Fuel	Climate zone	Solar status	Energy consumption		Daily temperature		Granular profile consumption	
			Pre-period	Post-period	Pre-period	Post-period	Pre-period	Post-period
Electric	100.0%	100.0%	99.7%	99.7%	99.7%	99.7%	99.7%	99.7%
Gas	100.0%	100.0%	99.8%	99.8%	99.8%	99.8%	99.8%	99.8%

We also replicated the PA's site-level results to a considerable extent. Table 4-2 summarizes our replication efforts by fuel and program across all three program years. As the values in the table indicate, the differences between the DNV and the PA reported per site and total energy savings estimates were no greater than 1.2% for both fuels, indicating the success of our replication of the PA M&V results.

Table 4-2. Comparison of DNV and PA M&V results by program and fuel

Fuel	Program	Absolute percent difference	
		Per site impact	Total energy savings
Electric	P4P - HomeIntel	0.0%	0.0%
	P4P - Comfortable Home Rebates	0.2%	0.2%
	P4P - Home Energy Rewards	0.5%	0.6%
Gas	P4P - HomeIntel	0.7%	0.9%
	P4P - Comfortable Home Rebates	0.0%	0.1%
	P4P - Home Energy Rewards	0.9%	1.2%

Program savings. Table 4-3 summarizes the first year claimed, PA M&V, and evaluated (validated) savings. It also provides evaluation realization rates (evaluated relative to the PA M&V savings, highlighted in brown in the table). Most evaluation realization rates for the P4P programs are close to 100% because DNV was able to replicate and validate the PA's M&V savings estimates for most participants. The table also includes the values of the PA M&V and evaluated savings relative to the final claimed savings (PA M&V and gross realization rates, respectively). Except for the PY2019 claimed savings, which the PA did not true-up, most final claimed electric savings were up to six times higher than the PA M&V and evaluated savings. Gas savings were similarly higher for two of the programs. The table also provides the percent savings (savings per site relative to average annual energy consumption) and relative precisions for the evaluated savings. We discuss the results for each program in detail below.

Table 4-3. Summary of P4P electric and gas gross savings, PY2019 to PY2021

Program	Program year	Electric sites	First year savings (kWh)							
			Claimed savings	PA M&V results	PA M&V realization rate	Evaluated savings	Evaluation realization rate	Gross realization rate	Percent savings	Relative precision
P4P - HomeIntel	2019	404	595,895	391,318	66%	391,318	100%	66%	7.5%	17%
	2020	818	643,945	561,852	87%	561,825	100%	87%	7.7%	15%
	2021	2,648	1,994,086	1,185,439	59%	1,185,409	100%	59%	6.7%	10%
P4P - Comfortable Home Rebates	2019	562	2,883	57,374	1990%	57,373	100%	1990%	1.1%	98%
	2020	2,053	638,753	861,411	135%	860,447	100%	135%	4.8%	14%
	2021	2,276	3,201,543	1,386,773	43%	1,383,443	100%	43%	7.2%	10%
	2019	484	38,085	102,994	270%	102,994	100%	270%	2.3%	53%

Program	Program year	Electric sites	First year savings (kWh)							
			Claimed savings	PA M&V results	PA M&V realization rate	Evaluated savings	Evaluation realization rate	Gross realization rate	Percent savings	Relative precision
P4P- Home Energy Rewards	2020	5,005	429,650	155,945	36%	155,597	100%	36%	0.5%	80%
	2021	1,779	673,671	108,948	16%	107,230	98%	16%	1.0%	112%

Program	Program year	Gas sites	First year savings (therm)							
			Claimed savings	PA M&V results	PA M&V realization rate	Evaluated savings	Evaluation realization rate	Gross realization rate	Percent savings	Relative precision
P4P - HomeIntel	2019	462	39,163	23,056	59%	23,025	100%	59%	6.1%	39%
	2020	789	79,960	14,479	18%	15,176	105%	19%	3.4%	37%
	2021	2,516	39,734	40,669	102%	40,720	100%	102%	3.5%	21%
P4P - Comfortable Home Rebates	2019	43	305	981	322%	994	101%	326%	3.9%	164%
	2020	1,346	24,052	36,531	152%	36,358	100%	151%	6.4%	18%
	2021	2,002	54,751	83,720	153%	83,818	100%	153%	9.1%	12%
P4P- Home Energy Rewards	2019	535	4,768	9,021	189%	8,847	98%	186%	3.3%	72%
	2020	4,024	96,534	29,344	30%	28,797	98%	30%	1.6%	32%
	2021	1,581	114,394	21,642	19%	21,667	100%	19%	3.5%	21%

HomeIntel. The participant counts for HomeIntel reflect the ongoing growth of the program. Program electric savings have increased over the three program years, and PY2021 program gas savings are well above the previous two years. The increasing population counts drove the increasing aggregate program-level savings as the underlying average electric and gas consumption per participant dropped by almost 50% relative to PY2019 levels.[50]

The electric percent savings (which reflect average savings per site relative to average annual consumption) were stable across the three years at approximately 7%. Not surprisingly, given the consistently high percent savings, the electric estimates for all years meet the population NMEC goal of 25% relative precision at 90% confidence.[51]

HomeIntel gas percent savings, in contrast, decreased by almost half in PY2020 and PY2021. With lower gas percent savings but a much larger program population in PY2021 than in the earlier years, the relative precision of the PY2021 estimate also meets the 90/25 population NMEC goal. The lower gas savings percentages of the latter two years are likely a better indicator of future gas savings for the program. If the program size of PY2021 can be maintained or increased, the strong relative precision results should continue. Overall, it is unlikely average participant consumption will continue to decrease, so PY2021 savings magnitudes are likely a reasonable indication of future program performance.

Comfortable Home Rebates. The Comfortable Home Rebate program population increased four-fold between PY2019 and PY2021, while program savings increased even more dramatically. Unlike HomeIntel, Comfortable Home Rebate average participant consumption remained roughly consistent across the three years, while savings as a percentage of consumption increased considerably each year from PY2019 to PY2021 for both electricity and gas. For PY2020 and PY2021, both gas and electric savings estimates had relative precisions substantially better than the 90/25 target. It seems unlikely the upward trend in savings as a percentage of consumption can continue. However, the Comfortable Home Rebate program is already successfully producing notable savings.

Home Energy Rewards. The Home Energy Rewards program was the biggest of the P4P programs in PY2020 but fell back to the smallest in PY2021. The surge in PY2020 occurred because the Home Energy rewards program, with its energy efficiency kits, was well-positioned to continue and even accelerate implementation during COVID shutdowns. Program

[50] This is consistent with the discussion in section 4.2.1, which indicates that HomeIntel widened its target households after the first year to include participants with more modest consumption levels.

[51] Population NMEC programs aspire to a precision of 25% at 90% confidence. Relative precision is a function of the magnitude and variability of the savings and the number of participants.

savings followed a similar pattern complicated by substantial changes in average participant consumption and savings as a percentage of consumption. The reduction in average participant consumption was not as dramatic as for HomeIntel but put a downward trend on savings potential. While Home Energy Rewards, with its focus on small measures, is expected to have lower savings than the other two P4P programs, the observed variability of savings, ranging from 0.5% to 2.3% of consumption for electricity and 1.6% to 3.5% for gas, was notable. The lowest percent savings was for cohorts that started during 2020, the first year of the pandemic. While the two other P4P programs did not demonstrate a COVID-related dip in performance, the dramatic increase in HER program enrollment could explain the HER results in PY2020. The poor relative precision results reflect savings that are small percentages of participant consumption. For a program producing savings this small, meeting 90/25 benchmarks will always be challenging, except with large populations.

P4P validation. While the range of potential validations is unlimited, examining how a study defines and addresses exclusions is critical, as this has the most potential to affect the results. Standard billing analysis practice tracks "attrition," the customers removed from the analysis due to lack of data and other exclusion criteria. Unless the exclusion rules cause the eligible analysis subset to be unrepresentative of the wider population, standard billing analysis applies estimates of impacts from this group to all customers. It is common for final analysis datasets to represent 70% of the starting program participants. For some populations, such as low-income customers, attrition can increase to 50%. The challenge in a population NMEC context is to make sure that exclusion rules are developed and applied in a way that is reasonable and will not unduly bias the final estimate. The eligibility requirement that all participants have a year of data to characterize the baseline period for population NMEC should reduce attrition relative to a standard after-the-fact billing analysis without such eligibility requirements.

The PA applied the same exclusion rules to all three P4P programs. The updated May 2023 M&V Plan for P4P programs provided a clear overview of general exclusions. It laid out the official eligibility requirements, which primarily ensure the availability of billing data for the evaluation period (12 months before participation and 24 months post) for sites where program activity occurs. In addition, the plan indicated that participation in other programs affects eligibility for P4P program participation. More importantly, the final PA M&V code documented the complicated exclusion process. The code goes beyond the M&V plan in identifying conditions that make a performance-based estimate of savings unreasonable and the type and timing of the alternative assigned savings those sites receive. We can simplify the exclusions into two baseline and three reporting (or performance) period categories. Each participant falls into one of the six permutations possible between the two types of exclusions.

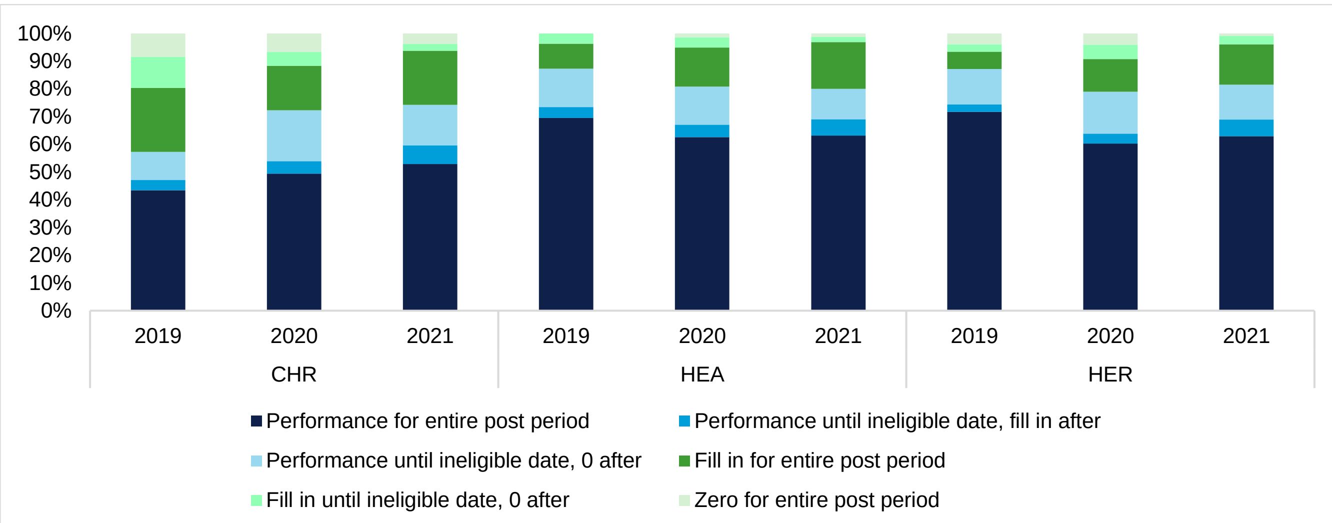
The baseline exclusions effectively determine whether the data is good enough to produce a reasonable performance-based estimate of savings for the site for at least part of the reporting period. The alternatives to performance-based savings estimates are assigned savings equal to the most relevant identified group cohort average savings or zero. The reporting period exclusions determine if there is a date when the basis for savings changes. If the reporting period starts with performance-based savings, occurrences at the customer site, such as moving, participation in another program, or change in solar or EV status, can turn those to either assigned savings or zero. Similarly, if the post-installation period starts with assigned savings, occurrences at the customer site can turn those to zero.

Figure 4-2 summarizes the effect of the exclusion process on electric savings by program and year.[52] The first three categories from the bottom of each bar use performance-based savings estimates for at least part of the post-installation period. The percentage of customers with at least some portion of savings informed by performance-based estimates ranges from 57% to 87%, with PY2021 levels all above 74%. The share of customers for whom savings are fully performance-based is more than 50% for the Comfortable Home Rebate program and 60% for the other two programs. These are consistent with general billing analysis results that reflect savings estimates based on 50% to 70% of participant

52 The exclusion process for gas indicates an even higher percentage of performance-based savings.

data. The subset of customers who move from performance-based savings to zero savings mostly ranges from 11% to 15%. This subset includes customers who move out (the largest category), opt-out, add solar, or participate in another program. The assignment of zero savings to these customers is a more conservative choice than in typical billing analysis practice. When evaluating programs with measure installations, it is common to assume that savings for the sites where customers move out will be, on average, the same as those who remain. Behavioral programs generally set savings to zero after move-outs, but even this is a conservative practice. For example, a recent analysis demonstrated that roughly 50% of Home Energy Report program savings remain with the house for some time, regardless of the occupant.[53] Given the varying mix of measure-based and behavioral savings for the P4P programs, it is difficult to determine the ideal level of assigned savings. The current exclusion practice errs on the side of caution, which seems more appropriate for NMEC programs.

Figure 4-2. Exclusion outcomes by program and year



Given the high percentages of customers who receive performance-based savings, the share of those who receive non-zero assigned savings for the entire post-installation period is relatively modest. The analysis assigns savings from a progressively identified group cohort to this group and all customers with partial post-installation period savings. Where possible, average savings are defined by the combination of solar status, electrification, program, and installation month cohort. Savings are assigned from the identified group cohort at this level if the group with performance-based estimates is sufficiently populated to provide a stable estimate. The analysis removes each characteristics level successively to reach a stable average with which to fill the site-level savings. This process is reasonable, though it might be worthwhile to understand the basis of the priorities of the grouping characteristics. While solar status and electrification are important categories, assigning a single savings value based on averages across programs as different as these three programs is difficult to understand. The energy savings of future programs could be even more variable. Additionally, the installation month cohort seems less important than the other three characteristics. There may be sound reasons for this ordering, but it is necessary to understand these to offer alternatives.

One baseline period exclusion that can lead to assigned savings is a lack of baseline data. For an NMEC program with clearly stated eligibility requirements regarding the availability of baseline data, this sort of exclusion should only occur under rare circumstances. The occurrence of this exclusion proves to be appropriately rare. Across the Comfortable Home Rebates and HomeIntel programs in PY2021, fewer than 1% of households were flagged for unavailable data. The Home Energy Rewards program flagged 3% of sites for inadequate data. If percentages of customers with insufficient data were to

[53] Brandon, Alec, et al. "Do The Effects of Nudges Persist? Theory and Evidence from 38 Natural Field Experiments." August 5, 2022. https://pages.jh.edu/alec/movers.pdf.

become a problem, the obvious solution would be to zero out savings for those customers rather than assigning identified group savings, as in the current case. The other common baseline exclusion, having a relatively uninformative baseline model with CVRMSE greater than 1, affected 10% to 16% of households in PY2021. In the context of a residential P4P program, this level of poor CVRMSE models is reasonable. In this context, it would also be burdensome for the programs to assess CVRMSE before allowing participation. Given this, the assignment of identified group average savings for these participants seems appropriate.

4.1.1.2 OBF gross savings validation

The PA used two different M&V methods to estimate PY2020 and PY2021 OBF program savings.

- For PY2020, it calculated savings using a difference-in-difference approach with granular profiles.
- For PY2021, it calculated savings using a synthetic control approach with granular profiles.

Replication process. DNV based its replication of the PA OBF M&V results using the provided methods document and data. As detailed in the methods section, we used the information we received to generate site-level savings that we aggregated to determine cohort-level savings.

Figure 4-3 summarizes the process DNV used to generate the cohort-level savings. We joined raw project characteristics, interval usage, associated granular profile usage, and weather data and cleaned the data to produce the prepared data. We then performed site-level regressions to generate daily site-level avoided energy use and savings. We rolled up site-level daily to annual savings that we aggregated to produce the cohort-level savings. We excluded sites with poor model fits from the cohort-level savings. The methods document provided by the PA included detailed descriptions of a process to assign savings to these sites. However, the PA did not include the assigned savings in the final PY2020 reported savings. Thus, our evaluated savings also excluded these savings. For PY2021, the PA M&V process produced savings that included customers with assigned savings.

Figure 4-3. Steps used in PY2020-PY2021 OBF Results

Electric program savings. Table 4-4 provides the claimed electric savings, the PA M&V results and realization rates relative to claimed savings, the evaluated gross savings and evaluation realization rates relative to PA M&V savings, the overall gross realization rates (evaluated relative to claimed savings), percent savings, and precision estimates for the OBF program. The PA M&V results reflect different methods in PY2020 and PY2021. While the OBF program did not change between the two program years, many aspects of the M&V process differed between the two years. As a result, comparisons across the years need to be done with care. For example, the realization rates differed primarily due to changes in the savings claim and M&V processes rather than how the programs operated.

Table 4-4 Summary of OBF electric gross savings, PY2020 - PY2021

Program year	Number of participants	First year savings (kWh)							
		Claimed savings	PA M&V results	PA M&V realization rate	Evaluated savings	Evaluation realization rate	Gross realization rate	Percent savings	Relative precision
2020	322	2,436,946	2,436,946	100%	2,466,287	101%	101%	1.7%	238%
2021	392	10,206,131	13,774,580	135%	8,596,574	62%	84%	7.0%	86%

For PY2020, the PA trued-up claims based on an early M&V process and only claimed electric savings for sites without exclusions.[54] This is reflected in the 100% PA M&V realization rate. DNV replicated the electric savings estimates closely for sites without exclusions. However, we could not replicate the application of the exclusions or site-level savings estimates for customers that the M&V process excluded. Given the change in the M&V approach and the exclusion process for PY2021 and forward, further exploration of the 2020 exclusions offered limited potential for insight. As a result, rather than explore the sites with exclusions, we provide the replication of results only for those with claimed savings. The PY2020 electric savings reflected the non-weather-normalized, calendar year 2021 savings rather than the expected weather-normalized savings for the post-installation year for each site. Because the OBF program pre-dates the NMEC Rulebook and the decisions made did not lead to an increase in savings, we did not adjust for these dimensions in the analysis.

The PA PY2021 M&V results for OBF reflected a different M&V methodology. The PY2021 OBF methods were substantially more consistent with the P4P program M&V processes. The PA M&V results were higher than the original claimed savings by 35%. DNV replicated the PA's M&V process and PY2021 electric OBF savings. Unlike PY2020, the PY2021 results included results for almost all customers without and with exclusions. Similar to P4P, the M&V process assigned savings to customers with exclusions. However, because most sites had preliminary engineering savings estimates, assigned savings, when necessary were based on the engineering estimates adjusted by how well other projects by their implementer performed on non-excluded sites. Given the availability of engineering estimates for most projects, this is a sound alternative to savings applications when exclusions make a performance-based site-level estimate infeasible.

Electric savings validation. While the PY2020 electric evaluation realization rate is effectively 100%, the aggregate electric savings estimate is not statistically different from zero, meaning that from a statistical perspective, we cannot be confident the program saved any energy. The relative precision estimate of over 100% captures this.[55] We also estimated weather-normalized savings, consistent with NMEC Rulebook guidance. Although this increased the savings, it did not achieve statistical significance.

We calculate precision for population NMEC using the variation of the mean site-level savings across participants. The poor precision for PY2020 reflects wide variability in the savings estimates partly due to the wide variation in customer size. While there are rules to limit claims using population NMEC methods for customers representing more than a fixed percentage of program baseline consumption (4% in PA M&V analysis code), this restriction may not be sufficient to support statistically significant results.

A small number of large, negative savings estimates were the cause of lower-than-forecasted program savings and the poor precision for PY2020 OBF. The final claimed PY2020 OBF electric savings represented only 8% of the preliminary savings filed in the initial claim. The final true-up produced a final claim that reflected a drop in savings from roughly 20% of consumption to just 2%. This reduction, caused by the performance-based savings estimate, was likely more than just a correction of poor preliminary savings estimates. Many of the largest sites had negative savings estimates, which was likely due to uncontrolled-for non-routine events (NREs).[56] While NREs that affect savings in both directions will exist throughout the population, the low final claimed savings and lack of precision appear to have been caused by an unfortunate concentration of NREs increasing consumption among large customers. The NMEC Rulebook explicitly states that population NMEC requires a substantially homogeneous population. Clear bounds on the range of customer sizes allowable within a single NMEC program have not yet been set. These results illustrate the risks of not doing so.

[54] This is made clear by comparing the M&V results, which provide savings based on a performance basis separate from those that are backfilled using savings based on, in this case, preliminary engineering-based savings de-rated by a 0.95 NTG rate.

[55] Relative precision is a way to express how statistically confident we are that an answer is within a range. A 90/10 relative precision indicates that we are 90% confident that the true result falls within 10% on either side of the estimate. In this case, we can only be 90% confident that the true value is within 238% of our finding, which means it could be zero. NMEC has multiple eligibility requirements explicitly designed to avoid this outcome.

[56] Non-routine events are energy use changes not caused by changes in the explanatory variables used in the prediction models or by the energy efficiency interventions.

In PY2021, the preliminary engineering-based savings estimates, which were the basis of initial claims, were again substantially lowered with the true-up process. Savings as a percentage of consumption dropped from roughly 25% to just 7%. Again, this was due to a combination of unrealistically high preliminary estimates and negative savings that diminished savings. In this case, the preliminary savings estimates were even higher than PY2020 and appear to explain the better part of the difference.

For the PY2021 evaluation, the application of exclusions is an important focus of the validation process because this process is more representative of how exclusions will be applied going forward. The M&V process removed altogether 6 sites whose energy consumption exceeded 4% of the total program baseline consumption. It also excluded 33 sites that were flagged as solar or had poor baseline period models and 9 sites with insufficient consumption data.[57] These exclusions represent a relatively low percentage of the population. The assignment of savings to excluded sites based on their adjusted engineering savings values is roughly consistent with the P4P method of savings assignment.

During the replication effort, DNV discovered two additional sites flagged for exclusion without documentation. These sites were simply identified by Customer ID in the code and were not identified through any of the documented channels. These two sites had negative savings. Including their performance-based savings estimates dropped the program's realization rate from 100% to 62% and worsened the relative precision from 24% to 86%. While removing these sites produced the sought-after precision level, this kind of ad hoc analysis raises concerns with the concept of embedded evaluation. In a less time-constrained context, the PA M&V process could have reached a different decision, or at least it could have highlighted and justified the decision to replace the performance-based savings with adjusted engineering savings. Instead, this provides a clear example of why replication and validation of NMEC programs are required.

Gas program savings. Table 4-5 provides the same set of evaluation results for OBF gas as Table 4-4 provided for electric. Again, the PA M&V results reflect different methods in PY2020 and PY2021. The gas claimed savings were not trued-up in CEDARS for either program year. The PA M&V results increased both estimates substantially.

Table 4-5. Summary of OBF gas gross savings, PY2020 to PY2021

Program year	Number of participants	First Year Savings (therm)							
		Claimed savings	PA M&V results	PA M&V realization rate	Evaluated savings	Evaluation realization rate	Gross realization rate	Percent savings	Relative precision
2020	28	5,317	148,573	2794%	51,741	35%	973%	2.6%	210%
2021	28	123,282	538,942	437%	423,454	79%	343%	80.7%	66%

DNV's ability to replicate the PY2020 gas results was limited. It was clear from the data provided that, for example, the regressions for all sites used a single granular profile as the comparator for all participants rather than being matched to industry-type and region-specific profiles. As a result of this and other issues, our savings estimates and the identification of exclusions were substantially different from the M&V results. DNV's estimates reflect our attempt to estimate savings using the correct granular profiles and appropriately applying exclusion factors from the M&V plan at the time. The PY2021 gas results were fully replicable but some exclusions were not applied consistently. The PY2020 savings estimate was not statistically different than zero at 90% confidence while the PY2021 savings estimate was statistically different than zero but did not meet the intended precision benchmark.

The substantially lower number of gas than electric participants reflects that most OBF projects are lighting projects with only electric savings. It also highlights the absence of a requirement for gas interactive effects in the savings claim either via performance-based methods or deemed for population NMEC. There is an ongoing discussion in the context of site NMEC regarding the need for guidance on dealing with interactive effects.

[57] The solar flag was not documented, so it was not clear whether it reflected the presence of solar or the addition of solar or both.

Gas savings validation. While DNV could replicate the PY2021 gas results, we found the application of exclusion inconsistent. The lack of an updated OBF M&V plan, which only became available after DNV's evaluation and most of the reporting was completed, and the lack of clarity over the actual application of exclusions noted in the existing M&V plan complicated the issue. Over half of the sites were missing pre- or post-installation models but were still assigned savings. These assigned savings represented over half of the claimed savings. In addition, a potentially important exclusion factor that was part of the original OBF M&V plan was not applied. That exclusion required that savings exceeding 50%, positive or negative, receive assigned savings instead of their calculated savings. Finally, the PA M&V process did not remove any sites due to size-related reasons, likely because of the relatively small size of the population.[58] While this change in exclusions for gas may be understandable due to the small size of the population, this allowed a single site to represent 39% of the baseline period consumption and 44% of the expected savings.

After replicating the PA M&V results, DNV re-calculated the savings with a modest change to exclusions. The high fraction of sites without baseline data was concerning. Generally, sites with a poor baseline model should still provide baseline consumption. The implication was that data was insufficient for all of those sites. Ultimately, we removed six sites that had neither site-level data nor preliminary savings estimates altogether. These sites did not have engineering estimates with which to assign savings and received program average savings in the PA M&V calculation. The result has an evaluation realization rate of 79%. The 66% precision at 90% confidence is different from zero but not near the target precision of 90/25.

The PY2021 percent savings of 81% (the ratio of average savings to average consumption) requires further explanation. Only 13 of the 22 sites included in this result have baseline data available to calculate average consumption, while all 22 have savings estimates after the savings assignment process. However, there is no other reasonable way to compute this percentage, and the unreasonably high percent savings illustrates the problems with the underlying data. Two sites, one with performance-based savings that are almost 100% of baseline consumption[59] and a second with assigned savings, represent over half of the final savings estimate.

The issues raised in this section are discussed further in section 4.2.

4.1.2 Program attribution and net savings

We provide our findings of program attribution for the P4P and OBF programs in this section. We provide details on how we developed the weights based on the achieved sample to expand results to the population in Appendix E.

4.1.2.1 Program attribution for P4P programs

HomeIntel is a behavioral program that educates participants about their energy load profile and consumption patterns and recommends actions to reduce their energy consumption. Since this is a behavioral program that does not incentivize any EE technology, we based the program's NTGR solely on the respondents' self-reported likelihood of taking the program-recommended action(s) in the absence of the program rather than on timing, efficiency, and quantity (TEQ) questions. The results in Table 4-6 indicate a high proportion of participants (55%) would have undertaken the actions without the program's recommendations. The portion of achieved savings attributable to the program is thus relatively low, with an NTGR of 45%.

[58] This exclusion still exists, partially, in the current evaluation code but gets dropped because it also checks for an electrification flag that does not exist in the data.

[59] Though there are indications in the code that an electrification flag exists, it was not present in these data. In addition to this extreme savings level, the site generated positive electric savings, indicating that electrification was unlikely to be the explanation.

Table 4-6. P4P PY2019-PY2021 program attribution - HomeIntel

Program name	Program ID	Measure name	Free-ridership	NTG	Program relative precision +/-
HomeIntel	PGE_RES_001B	Behavioral Intervention	55%	45%	5%

The CHR program offered participants rebates to retrofit their homes through an improvement pathway or to upgrade their HVAC system through a maintenance pathway. The program's rebates covered about 10% of retrofit project costs and over 90% of HVAC maintenance costs. Table 4-7 provides estimates of program attribution for the CHR program by measure, program pathway, and overall. Program attribution was higher for the HVAC maintenance pathway, with an NTGR of 93% than for the retrofit pathway, which achieved an NTGR of 86%. DNV estimated the CHR program's overall NTGR to be 89%. Timing was the biggest driver of attribution for this program, indicating that the program accelerated the timeline of purchasing and installing the EE measures offered by the program compared to non-program scenarios.

Table 4-7. P4P PY2019-PY2021 program attribution - Comfortable Home Rebates

Path	Measure name	Weight	NTG			Program relative precision +/-
			Measure	Path	Program	
Improvement	AC Replacement	83	83%	86%	89%	1%
	Attic Insulation	46	80%			
	Water Heating Replacement	4	96%			
	Air Sealing	46	84%			
	Deep-Buried Ducts	15	82%			
	Duct Replacement	87	84%			
	Heat Pump HVAC	6	78%			
	Heating Replacement	56	91%			
	Right-Sized Returns	18	100%			
	Smart Thermostat	146	88%			
Maintenance	Motor Replacement	7	95%	93%		
	Fan Controls	41	89%			
	Coil Cleaning	10	94%			
	RCA	31	95%			

The HER program provided free energy-saving kits and EE measures through an online marketplace. The EE kits consisted of LEDs, showerheads, and aerators, and the online marketplace offered the kit measures, smart power strips, smart thermostats, and a home energy monitor. Table 4-8 provides estimates of program attribution for the HER program. It shows that the online marketplace measures had a higher attribution, with an NTGR of 86%, than kit measures, with an NTGR of 66%. The overall NTGR for the HER program was 73%. Smart thermostats were the most purchased measure through the online marketplace in the sample. Since we found the program smart thermostats to have low free-ridership, their prevalence helped drive up the overall high NTGR for the program.[60]

Table 4-8. P4P PY2019-PY2021 program attribution - Home Energy Rewards

Path	Measure name	Weight	NTG			Program relative precision +/-
			Measure	Path	Program	
Kits	EE Kits	1,182	66%	66%	73%	2%
Online Marketplace	Aerator	20	82%	84%		
	HAN	24	91%			

[60] Discounts for smart thermostats through the online marketplace were as high as 60%. In addition, the program offered rebates for these devices, which allowed most participants to obtain them for free. This incentive level likely reduced the acquisition barrier sufficiently and led to high attribution for the measure.

Path	Measure name	Weight	NTG			Program relative precision +/-
			Measure	Path	Program	
	LED Lighting	163	71%			
	Showerhead	6	45%			
	Smart Thermostat	353	91%			

4.1.2.2 Program attribution for OBF programs

Representativeness of the OBF respondent sample. DNV used a representative sample drawn from projects with 2020 and 2021 installations to determine OBF program attribution. Table 4-9 summarizes the claimed savings and the number of projects for the population and the respondent sample. The sample used to estimate program attribution represented 23% of all claimed savings and 12% of all projects.

Table 4-9. OBF PY2020-PY2021 population and sample claimed savings and projects by project type

Project type	Claimed savings (MMBtu)			Number of projects		
	Population	Sample	Percent population	Population	Sample	Percent population
Lighting	151,559	32,511	21%	698	78	11%
Both	56,480	9,066	16%	56	9	16%
Non-Lighting	51,172	17,722	35%	79	12	15%
All	259,211	59,299	23%	833	99	12%

For context, Table 4-10 summarizes the percent of claimed savings and projects covered by the DNV and the two prior studies that provided estimates of OBF program attribution. At first glance, the DNV and PG&E 2018-19 samples represent a comparable percent of claimed savings. However, the PG&E study included one large site, without which the sample the PG&E study used covered only 4% of claimed savings.

Table 4-10. Comparison of OBF PY2020-PY2021 population and samples with past studies

Group	OBF-AP 2020-2021	OBF-AP 2018-2019	PG&E 2018-2020
Number of projects			
Population	833	202	163
Sample	99	18	23
% Population surveyed	12%	9%	14%
Reported gross savings (MMBtu)			
Total savings	259,211	131,235	117,773
Sample savings	59,299	9,818	34,686
% Total savings	23%	7%	29%

Estimate of OBF Free-ridership and Influence. DNV asked participants to indicate how likely they were to seek alternative financing to complete their projects without OBF. As stated in section 3.2.2.2, we used the responses to this question to determine how we estimated program attribution. Table 4-11 summarizes participant responses to this question. The results indicated that 4% of respondents would have likely sought loans (in the form of credit cards with interest, loans, and eco-green funds) to complete projects without OBF. Similarly, 6% of ODC's OBF-AP 2018-2019 participants indicated they would have likely sought and secured alternative financing without OBF.

Table 4-11. Methods likely used in the absence of PY2020-PY2021 OBF

Likely payment method	Number of responses	Weighted responses	Percent response
Cash	73	597	71%
Credit card paid in full	2	8	1%
Interest-bearing loans (credit card, loan, eco-green financing)	4	35	4%

Likely payment method	Number of responses	Weighted responses	Percent response
Grant	20	165	20%
None	3	31	4%

For the 4% of respondents who reported they would have likely sought alternative financing, we used their response to the likelihood of completing the project without OBF to determine their program attribution. The responses from these participants indicated that 54% would have done the project without OBF.[61]

We used responses to general financing counterfactual scenarios to determine program attribution for the remaining 96% of respondents. For these respondents, we measured program attribution using the following formula: P * T * E * Q, where P is a 0/1 probability the project would have happened without financing. We used participant responses to the likelihood they would have done the project without financing to determine this probability. Respondents who indicated they were "very unlikely" to complete the project without financing received a value of P = 0 and were zero free riders. Approximately 17% of respondents fell in this category. For the remaining respondents, who received a value of P = 1, we used their responses to the TEQ questions to determine the free-ridership score. The overall NTGR reflected the combination of the responses from these two general respondent types. Table 4-12 provides NTG values for PY2020 and PY2021 OBF population NMEC programs and measures.

Table 4-12. OBF PY2020-PY2021 attribution by measure and overall

Program name	Program ID	Measure	Weight	NTG	Program relative precision +/-
OBF	PGE210911	HVAC	4.3	38%	38%
		Lighting	9.4	57%	9%
		Refrigeration	6.2	45%	61%
		Overall		56%	9%

As the table indicates, the overall NTGR estimate for OBF projects with 2020 and 2021 installations was 56%. This NTGR is lower than those reported in previous evaluations. The PY2018-2019 ODC evaluation estimated the NTGR of the OBF-AP program to be 69%, while the PG&E study based on installations from the same period estimated the NTGR for this program to be 94%. However, the current evaluation relies on a more diverse and robust sample with 99 measure-specific responses (representing HVAC, lighting, and refrigeration) and a relative precision of ±9%. This contrasts with the limited samples used in the previous evaluations. In the PG&E 2018-2019 case, the program NTGR ratio reflects a program attribution value of 100% from a single project, accounting for over 80% of the program savings in the survey sample. Without this project, the average NTGR of the remaining projects included in the PG&E 2018-2019 would have been more in line with other recent evaluation results, including the current study.

Like the PG&E 2018-2019 study, most survey respondents (63%) in DNV's evaluation indicated they had planned their project before learning about OBF. A substantial proportion of these (81%) had also included the project in the organization's capital budget before learning about OBF. While inclusion in organization capital budgets does not guarantee projects' implementation due to other competing projects that need financing, this finding in the current study tends to corroborate the relatively high level of free-ridership DNV estimated.

Drivers of OBF Free-ridership. To put DNV's estimate of program attribution into context, particularly relative to the PY2018-PY2019 OBF-AP findings reported by ODC, we classified respondents into three free-ridership categories. These included non-free riders, partial free riders, and full-free riders. As Table 4-13 indicates, respondents whose free-ridership score was no higher than 20% were non-free riders, those whose free-ridership score was from 20% to 70% were partial

[61] By comparison, 52% of those unlikely to seek alternative financing reported they would have completed the project without any financing.

free riders, and respondents who scored above 70% were full-free riders. Unlike the findings in the PY2018-2019 study, the DNV study indicated a lower proportion of respondents were non-free-riders (27% compared to 44%) while a higher percentage were full-free-riders (21% compared to 11%). The table summarizes the breakdown of the free-ridership scores into timing, quantity, and efficiency.[62] As the results indicate, OBF accelerated the timing of the projects for non-free-riders and encouraged partial free-riders to install more efficient technologies. Full free-riders would have undertaken projects with the same scope and efficiency at about the same time without the program.

Table 4-13. OBF PY2020-PY2021 free-ridership scores by timing, efficiency, and quantity scores

Free-ridership category	OBF free-ridership score range	Measure	Number of responses	Average timing score	Average efficiency score	Average quantity score
Non-free riders	0%-20%	Lighting	23	57%	61%	86%
		Non-lighting	3	67%	90%	
		All	26	58%	64%	86%
Partial free riders	20%-70%	Lighting	43	98%	61%	85%
		Non-lighting	8	100%	59%	
		All	51	98%	60%	85%
Full free riders	70%-100%	Lighting	16	97%	100%	89%
		Non-lighting	5	100%	95%	
		All	21	98%	99%	89%

Based on information from the table, both lighting and non-lighting projects had equal proportions (about 50%) of partial free riders, with a comparable percentage of respondents indicating that OBF enabled them to choose more efficient lighting and non-lighting measures. Lighting projects had a higher proportion of non-free riders (about 30%) than non-lighting projects (about 20%), with more respondents with lighting projects indicating that OBF helped them accelerate the installation of program measures.

Program influence by participant characteristics. We also sought to understand if program attribution differed by customer characteristics, including the size of OBF loans, organization category, and project type. The average and median loan amounts for the 2020 and 2021 OBF installations were approximately $100,000 and $37,000, and the maximum loan amount was about $3.8 million. Since the median loan among the respondent sample was higher, we examined program attribution for loans up to $50,000 and above $50,000.

We also classified participants into large commercial, industrial, and agricultural; institutional (such as public and education); and small and medium commercial sectors. We examined program attribution for the large commercial, industrial, and agricultural, and other organization categories.[63] We also considered program attribution for lighting and non-lighting projects.

As Table 4-14 indicates, program attribution was higher for lighting than non-lighting projects, smaller loans below $50,000 than larger loans, and for small and medium businesses and institutional sector participants than large commercial, industrial, and agricultural participants. Most non-lighting projects (92%) used loans exceeding $50,000. Thus, the lower attribution for projects with loans over $50,000 partly reflects the greater prevalence of non-lighting projects that used large loans. Large commercial, industrial, and agricultural sector participants had lower attribution than participants in other sectors because they undertook mostly non-lighting projects, which used loans over $50,000.

[62] The timing score measures the extent to which the program motivated participants to complete the project earlier, the efficiency score measures the degree to which the program encouraged participants to adopt a more efficient version of the installed technologies, and the quantity score measures the extent to which the program motivated participants to undertake a project with more scope.

[63] We determined NTGRs by all three organization types but found attribution to be similar between institutional participants and small and medium businesses. Thus, we reported NTGR values for these two sectors together.

Table 4-14. NTGRs by project type, loan size, and sector

Project type		Loan size		Sector	
Lighting	Non-lighting	$0 - $50,000	> $50,000	Institutional and Small/medium businesses	Large commercial, industrial and agricultural
58%	44%	63%	52%	57%	53%

Uncertainty in OBF program attribution. As both the Cadmus and ODC studies indicate, the influence of OBF-AP programs on savings is still uncertain and likely to be variable. The 2018-2019 Cadmus study found a high level of program attribution with an NTGR of 94% for OBF-AP, while the ODC PY2018-2019 study found an NTGR of 69% for the program. The Cadmus NTGR primarily reflected the 100% program attribution of a large single project, which accounted for over 80% of the program savings in the survey sample.

While the approaches used in the two studies were not the same, the authors of both studies acknowledged the level of uncertainty in attribution for programs that offer only on-bill financing and the need for additional studies. DNV's findings indicate OBF program attribution to be relatively lower at 56% and support the observation from the two studies regarding the variability in estimates of the influence of such programs. It is also possible that the COVID-19 pandemic had some impacts on the participant mix of the more recent programs.[64]

4.1.3 Total savings

As indicated in section 3.2.1, the PA final claimed savings reflect true-ups based on metered M&V results filed with the CPUC and provided in CEDARS.[65] The PA revised its M&V results, to account primarily for the effect of the pandemic.[66] DNV's evaluated results are based on the replication and validation of the revised PA M&V savings. We compared the evaluated values to the final claimed savings to generate gross realization rates.

Table 4-15 provides the final electric claimed savings (final claimed gross savings), the evaluated gross savings, and the gross realization rates for each population NMEC program. The evaluated gross savings are adjusted to reflect the influence of the programs using net-to-gross ratios (NTGRs) and generate net savings. Our evaluation indicates that the population NMEC programs in combination caused 0.3, 2.5, and 6.7 million kWh of electric savings in PY2019, PY2020, and PY2021, respectively.

Table 4-15. Population NMEC electric claimed and evaluated savings, PY2019 – PY2021

Program	Program year	Final claimed gross savings (kWh)	Evaluated gross savings (kWh)	Gross realization rate	Evaluated NTGR	Evaluated net savings (kWh)
P4P - HomeIntel	2019	595,895	391,318	66%	45%	176,093
	2020	643,945	561,825	87%	45%	252,821
	2021	1,994,086	1,185,409	59%	45%	533,434
P4P -Comfortable Home Rebates	2019	2,883	57,373	1990%	89%	51,062
	2020	638,753	860,447	135%	89%	765,798
	2021	3,201,543	1,383,443	43%	89%	1,231,264
P4P- Home Energy Rewards	2019	38,085	102,994	270%	73%	75,186
	2020	429,650	155,597	36%	73%	113,586

64 For example, it is possible that more marginal businesses, which would benefit more from OBF than their more economically secure counterparts, would be less likely to participate in the program during the pandemic because they had more urgent needs to address than installing energy-efficient measures. If this scenario played out, the remaining participants would tend to have lower NTGRs because they had more opportunities for alternative financing than would otherwise be the case if the participant pool were more diverse.
65 PY2019 final claimed savings do not reflect true-ups because the PA did not file these.
66 The PA M&V savings for all the P4P programs diverge substantially from the final claimed values. The PA M&V P4P electric savings are up to six times lower than the final claimed values.

Program	Program year	Final claimed gross savings (kWh)	Evaluated gross savings (kWh)	Gross realization rate	Evaluated NTGR	Evaluated net savings (kWh)
	2021	673,671	107,230	16%	73%	78,278
On-Bill Financing Alternative Pathway	2020	2,436,946	2,466,287	101%	56%	1,381,121
	2021	10,206,131	8,596,574	84%	56%	4,814,081

Similarly, we compared the evaluated to the final claimed gas savings to generate gross realization for population NMEC gas programs.[67] Table 4-16 provides the final gas claimed savings (final claimed gross savings), the gross realization rate, and the gross evaluated savings for each population NMEC program. Evaluated gross savings are also adjusted to reflect the portion of the savings that can be attributed to the program using net-to-gross ratios (NTGRs). Our evaluation indicates that the evaluated population NMEC programs, in combination, caused approximately 18, 89, and 341 thousand therms of gas savings in PY2019, PY2020, and PY2021, respectively.

Table 4-16. Population NMEC claimed and evaluated gas savings, PY2019 – PY2021

Program	Program year	Final claimed gross savings (therm)	Evaluated gross savings (therm)	Gross realization rate	Evaluated NTGR	Evaluated net savings (therm)
P4P - HomeIntel	2019	39,163	23,025	59%	45%	10,361
	2020	79,960	15,176	19%	45%	6,829
	2021	39,734	28,797	72%	45%	12,959
P4P - Comfortable Home Rebates	2019	305	994	326%	89%	885
	2020	24,052	36,358	151%	89%	32,359
	2021	54,751	83,818	153%	89%	74,598
P4P - Home Energy Rewards	2019	4,768	8,847	186%	73%	6,458
	2020	96,534	28,797	30%	73%	21,022
	2021	114,394	21,667	19%	73%	15,817
On-Bill Financing Alternative Pathway	2020	5,317	51,741	973%	56%	28,975
	2021	123,282	423,454	343%	56%	237,135

4.2 Application of NMEC rules

As a part of the evaluation process, we reviewed the application of the NMEC rules by the four population NMEC programs in the following areas: customer segments targeted and eligibility requirements, M&V plans, incentive structures, and overall program management. Our methodology and data sources for this assessment are outlined in section 3.3 of this report. Table 4-17 below summarizes the design elements we evaluated, the corresponding guidelines outlined in the PA's Advice Letters, and our assessment of adherence to Rulebook requirements.

Table 4-17. Application of NMEC rules by population NMEC programs

Program design elements	NMEC Rulebook	PG&E Advice Letters	DNV's assessment
Eligibility criteria	Programs need to define participation eligibility criteria, such as: • Participation in other energy efficiency programs • The presence of electric vehicles, solar PV, storage • Tenant turnover	Participants must sign up through the implementer acknowledging participation and inability to participate in other incentive offerings.	All four programs provided eligibility criteria in their PIPs. However, these criteria did not always fully explain exclusions applied in the M&V process.

[67] Except for the Comfortable Home Rebates program, the PA M&V P4P gas savings are generally lower than (up to six times) the final claimed values.

Program design elements	NMEC Rulebook	PG&E Advice Letters	DNV's assessment
Permissible project types	Sites with similar: • levels of energy consumption • drivers of energy consumption • equipment holdings	N/A	The P4P programs, as with most residential programs, meet this requirement. The OBF program fails this criterion with respect to energy consumption and likely the other two areas.
Incentive payment structures	Ideally, 100% and at a minimum 50% of total PA payments should be made based on payable savings. PA payments may be made before final savings determinations and adjusted after the 12-month post-intervention period.	Aggregator payments will be determined based on gross energy savings through a PG&E-facilitated weather-normalized pre/post analysis.	The P4P programs conformed to the incentive structure laid out in the Rulebook. The participating customer repays the no-interest OBF loan regardless of performance.
Measures and interventions	Measures currently allowable through the deemed and custom energy efficiency programs. Other measures where program documentation and M&V Plan demonstrate that savings and EUL forecasts are reasonable.	Interventions based on customer needs that lead to persistent consumption reductions. Retrofit, such as HVAC and insulation, and BRO measures.	P4P measures included a successful combination of behavioral messaging and retrofit measures. OBF also included a successful combination of lighting and non-lighting measures (such as HVAC, refrigeration, etc.).
M&V plans	Implementers must develop and submit an M&V Plan as part of their bid that must include: • A description of the program target population and participant eligibility criteria. • Documentation of the expected costs, energy savings, and EUL of planned measures and intervention strategies. • Identification of the method(s) and calculation software that will be used to calculate savings.	Pre/post-intervention analysis of participant's metered energy consumption compared to a matched pair control group through a quasi-experimental design approach.	The P4P M&V plans were not well defined in any of the program documents available at the time of the evaluation. The DNV team gleaned some information on these plans through PA interviews and the M&V code set. Updated plans for both P4P and OBF were provided at the end of the evaluation cycle.
Program management	NMEC Rulebook mandates that programs track relevant program information. It does not provide a specific roadmap and leaves it to the discretion of program administrators to design their program tracking systems.	N/A	The population NMEC programs effectively track the most relevant information. Other findings in this report indicate that better tracking and documentation of information that feeds into eligibility exclusions should be the next focus as program management systems develop.

4.2.1 Customer segments and eligibility criteria

The NMEC Rulebook requires that programs include a description of the target population and eligibility criteria as part of their M&V plan. Programs set eligibility criteria to minimize the effect of non-program-related changes on M&V results since implementers are compensated based on metered savings. Programs establish these criteria to ensure that participants have the data required to support a performance-based savings estimate and do not engage in activities that could affect savings estimates via unaccounted-for NREs.

Certain factors, such as the addition of electric vehicles (EV), solar photovoltaic (PV) panels, battery storage, or participation in other energy efficiency programs, can impact the calculation of savings. The population NMEC programs that DNV evaluated included data requirements, conditions on participation in other programs, the treatment of EVs, PVs, and battery storage in their PIPs, recently updated M&V plans, and their M&V code.

The original and updated M&V plans indicate that customers who have inadequate pre-period data, change solar PV or EV status, and participate in more than one program are baseline ineligible. Data from such customers will be excluded from savings estimates. The plans also indicate that the PA will verify these conditions during customer acquisition to determine participation ineligibility. Thus, participation eligibility in a population NMEC program requires:

- Customers have sufficient data (12 months of baseline period energy consumption)
- Customers whose solar PV or EV status does not change during the baseline period
- Customers who do not participate in other EE programs

The plans also specify that the PA will check that customers have no plans to change the solar PV and EV status and remodel or expand their sites during the reporting period. If such changes occur during the reporting period, they affect how the PA calculates the savings.

While programs have eligibility criteria that determine participation, they also use similar criteria to determine payable savings. The programs apply these criteria based on baseline and reporting period changes. DNV used the updated P4P M&V plan that became available in May 2023 and the PA M&V code to summarize the exclusions that determine payable savings. The criteria based on these exclusions define the savings that PAs pay to program implementers, including those based on assigned amounts as we described in the gross savings impact section 4.1.1.

Table 4-18 summarizes the P4P program baseline and reporting period exclusions and their effect on payable savings. It is worth noting that for PY2021, the PA M&V followed updated exclusion rules for all four programs, which it applied retrospectively to prior years. Based on code review, the PA M&V process assigned savings to OBF projects for the whole reporting period if they have insufficient data, poor models (CVRMSE > 1), or an undefined solar flag.

Table 4-18. Exclusion rules for P4P program payable savings, PY2019 to PY2021

Baseline period exclusion	Implications	Reporting period exclusion	Implications
Insufficient data Poor models (CVRMSE > 1)	Assigned savings	None	Assigned savings
		Move-out / opt-out	Assigned savings, 0 savings after move out date
		Acquire solar and/or storage	Assigned savings, 0 savings after solar and/or storage install date
		Change in solar capacity or EV status	Assigned savings after status change
		Other EE program installation(s)	Assigned savings, 0 after other EE install date
None	Performance-based savings	None	Performance-based savings
		Move-out / opt-out	Performance-based savings, 0 savings after move out date
		Acquire solar and/or storage	Performance-based savings, 0 savings after solar and/or storage install date
		Change in solar capacity or EV status	Performance-based savings after status change
		Other EE program installation(s)	Performance-based savings, 0 after other EE install date

HomeIntel. In the initial stages, HomeIntel targeted PG&E customers with an average hourly baseload of more than 500W. However, early program results showed success for customers with baseloads under 500W. Based on that success, all residential PG&E customers are now eligible to participate in the HomeIntel program if they have not participated in any other PG&E-sponsored energy efficiency program. Participants must also have 12 months of baseline data to be eligible.

HomeIntel does allow customers with EV and PV installations to participate in the program. However, like the other two P4P programs, the PA changes the treatment of savings from such customers if their EV, PV, and battery status changes during the performance period.

Comfortable Home Rebates.The CHR program targeted customers who lived in single-family and 2-4-unit residential buildings within the California Energy Commission (CEC) Climate Zones 11, 12, and 13. The program was available to all areas of the PG&E service territory except those in the Bay Area Regional Energy Network (BayREN). PG&E used targeted emails to reach specific customers with high summer cooling and winter natural gas usage profiles.

The CHR program required customers to have one year of baseline data for at least one PG&E-supplied fuel (natural gas or electric). The program restricted customers from participating in other programs in the last year or installing similar measures within the past five years. The program did not preclude customers who had solar, EV, or battery storage from participating but the PA changed the treatment of savings from such customers if the status of any of these changed during the performance period.

Home Energy Rewards. The HER program targeted single-family residential customers. In addition to residing in single-family homes, participants needed a valid PG&E service ID to participate in the program. Moreover, participants could only have one electric (and one natural gas if applicable) meter at their homes. The program required participants to have 12 months of baseline data, which required they lived in the same house for at least 12 months before enrolling. The program also limited participants from enrolling in other energy efficiency programs. According to the PIP, customers who had EVs, solar, and storage could not participate in the program. However, the PA M&V code indicated otherwise. The PA changed the treatment of savings from these customers if the status of any of these technologies changed during the analysis period.

On-Bill Financing. The OBF program is exclusively available to non-residential customers, such as commercial and government entities. The program offers zero-interest loans that cover the entire cost of an energy efficiency project, including M&V costs. To be eligible for the program, PG&E applies two main criteria. First, the business must be a PG&E non-residential customer for at least 24 months. Second, PG&E reviews the utility payments made by the participant over the last 12 months to assess consistent payment history. A payment history screening determines the good credit standing required for loan approval. Tenant must also be on-site for 12 months to establish baseline conditions. The program limits customers to participate only in NMEC and customized retrofit-qualified incentives. According to the original M&V plan, customers whose solar or EV status changes during the program period, particularly during the baseline are not good candidates for population NMEC programs. The updated OBF M&V plan also clarifies that customers whose solar or EV status changes during the baseline period will be removed from savings estimates. The PY2021 PA M&V analysis flagged 7% of customers as "solar" without defining whether this meant solar was present or was added. It assigned these customers savings based on identified group cohorts for the performance period. This provides evidence that the program recognizes that such identifiable NREs need to be addressed explicitly. It does not make a distinction between baseline and performance period occurrence of these eligibility issues and the implications for savings assignment.

4.2.2 Permissible project types

The NMEC Rulebook recognizes that population NMEC methods are only appropriately applied to populations with reasonably homogeneous consumption and energy usage characteristics. The NMEC Rulebook requirement regarding population NMEC Permissible Project Types states, “sites can reasonably be expected to have similar types of equipment holdings, as well as drivers and levels of energy consumption.” (NMEC Rulebook 2.0, II.2.B.2). While this description fits the residential P4P program populations, it does not fit the OBF program population. For example, the non-residential customers in the PY2020 OBF population have pre-participation consumption ranging from greater than 13,000 MWh to less than 1 MWh and with preliminary savings ranging from 4,667 MWh to 2.5 MWh.[68]

The Rulebook reflects best practice billing analysis methodology that recognizes that homogeneous sites are required to produce acceptable program precision estimates. The extreme range of size and savings for OBF raises the more fundamental question of whether any comparison group or synthetic control approach can reasonably address the NRE challenge for such participants. Comparison groups in aggregate and granular profiles reflect the exogenous change that effectively represents NREs aggregated across a wider population. Can the site-specific potential NREs of a handful of large sites be adequately addressed by a granular profile that, by definition, reflects the overall average NREs of the related population? The technical details of this question require additional attention. However, it is safe to say that if this combination of approach and highly variable population were to fail, it would do so by demonstrating highly variable savings estimates that fail to achieve adequate precision.

The OBF program's aggregate savings estimates are not statistically different from zero, let alone meeting the higher standard of relative precision of 90/25.[69] This appears to be a direct result of unexpected negative site-level savings results (likely due to NREs) for large sites combined with the extreme variability in size and savings in the population. The OBF program has instituted size constraints over time. This evaluation's results indicate customer size is an important consideration, and those constraints in place at the time may not have been sufficiently stringent. There is a parallel but broader policy question regarding what levels of savings can be claimed using population NMEC versus site-NMEC methods. Site-level methods afford a greater scrutiny of NREs at individual sites. There is some yet-to-be-determined size range in a population NMEC program that will allow both the theoretical expectation that NREs are addressed and produce program results that consistently meet relative precision benchmarks.

4.2.3 Incentive payment structures

Incentive payment structures are critical for energy efficiency designs that use a P4P approach. P4P is an inherently market-based approach where program implementers are paid based on actual savings measured at the meter. In contrast, traditional energy efficiency designs focus on one-way incentives, where customers can retain the incentives whether the interventions lead or do not lead to savings. Therefore, ratepayers end up bearing the bulk of the performance risk. P4P programs shift the risk from the ratepayers to program implementers.

The NMEC Rulebook requires PAs to adhere to NMEC approaches to determine payable savings.[70] It recommends that 100% of payments for population NMEC programs reflect energy savings based on NMEC methods. At a minimum, 50% of program payments must reflect payable savings calculated using population NMEC methods. In cases where programs fail to meet the 50% minimum, PAs must submit program-level M&V plans in a pre-program advice letter for approval.

In the interviews conducted by DNV, the PA indicated that implementer compensations varied depending on program designs. For example, 50% of the payments for CHR, which provided customer incentives, reflected customer incentives, and 50% performance-based payments. On the other hand, the PA compensated HomeIntel and HER solely based on metered savings. In all these cases, the PA used the realized savings to pay implementers and mitigate the risk of overpayment. The PA staff also indicated that clawback rules were in place in case of overpayment. This scenario could occur when programs receive cash advances at program start-up. If final metered savings reflect an overpayment, clawbacks are applied.

68 Three sites had preliminary savings expectations greater than their baseline consumption.

69 90/25 is a common criterion used in energy efficiency evaluation, requiring that the research achieves 90% confidence that an estimated metric, such as estimated savings, falls within 25% of the true value to provide a statistically valid outcome.

70 In contrast, claimable savings are the savings PAs report to the CPUC before a formal evaluation.

OBF followed the recommendation of the PG&E Advice Letter, which indicated that aggregators (that provide energy-efficient products directly or through contractors) are paid a fixed rate per therm and kWh for the measured savings. The OBF design avoids the risk of one-way incentives, where ratepayers bear the project's cost if the projected savings do not materialize, by requiring customers to pay the entire loan amount through their bill. If customers can't repay the loan, they risk facing service disruption.

4.2.4 Measure and intervention selection

The Rulebook requires that population NMEC programs can offer existing deemed and custom program measures or measures with reasonable savings and EUL forecasts. As a part of this evaluation, we reviewed the measure and intervention selections provided in the PIPs to see if the four programs conformed to the guidelines outlined in the NMEC Rulebook.

The measures and interventions implemented by the programs differed depending on their target population. Table 4-19 summarizes the interventions offered by the programs. OBF has financing options available for program participants, HomeIntel and HER offer energy assessments and personalized educational advice to assist customers in making informed choices regarding their energy consumption, and HER and CHR offer EE measures to help participants reduce their energy consumption. All four programs selected measures that delivered savings and met the Rulebook requirements.

Table 4-19. Summary of population NMEC PY2019-PY2021 program measures

Programs	Saving measures	Non-savings/behavioral measures	Financing
HomeIntel	Education about energy use and recommendations for ways to reduce energy use	No- to low-cost measures and customized recommendations e.g., reducing run time, unplugging unused appliances	No
CHR	Space conditioning and water heaters, attic insulation, air sealing, duct replacement and sealing, fans, smart thermostats	None	No
HER	Low-cost kits and energy-efficiency products offered through a discounted online marketplace. Rebate for smart thermostats.	Online assessment of energy use, and customized tips and recommendations	No
OBF	Tier 1: Do not require any change in the building's footprint, e.g., HVAC maintenance and tune-up, LED lighting, thermostats, and refrigeration system upgrades Tier 2: Require alterations to the building, e.g., upgrading HVAC, improving the building envelope, and installing advanced lighting controls	OBF aspires to move away from a widget-based approach and include behavioral and operations and maintenance (O&M) savings.	Yes

HomeIntel. HomeIntel assisted residential customers in understanding their home's energy usage and opportunities for energy savings through a Smart Audit. The Smart Audit is a web-based application that uses algorithms to identify energy efficiency measure recommendations based on electric and gas usage data. Participants also had the option to work with an Energy Coach to reduce energy waste and install recommended energy savings measures. Most program-recommended measures and interventions were low- or no-cost, such as smart strips, timers, reducing run times, unplugging unused appliances, or other actions that reduce high plug load use. HomeIntel's program interventions have resulted in measurable savings and conform with the NMEC Rulebook requirements.

Comfortable Home Rebates. The CHR program focused on retrofit measures for residential customers. Specific measures included heat pump space conditioning and water heaters, attic insulation, air sealing, duct replacement and sealing, whole house fans, and smart thermostats. The CHR program aligned with the measure selection outlined in the PG&E Advice Letter and the NMEC Rulebook qualifying measures, which include retrofits for HVAC and insulation.

Home Energy Rewards. The HER program provided single-family and 2-4 multifamily unit residential customers with an initial online energy assessment to identify energy-saving opportunities. It also offered participants the option to get a mail-in kit or purchase discounted energy-efficient products from the online marketplace. The kit contained two showerheads, a kitchen aerator, two bath aerators, and four LED bulbs. Participants could purchase smart thermostats (ECOBEE and Nest) and other small energy-efficiency measures from the discounted online market. The interventions of this program delivered modest but measurable savings and aligned with the Rulebook requirements.

On-Bill Financing. The OBF program offered zero-interest loans to eligible non-residential customers who wanted to participate in high-quality energy efficiency projects. A wide range of measures were available through a project-based process supported by contractors, ensuring that customers had multiple options and projects aligned to customer needs. The program categorized projects into two tiers. Tier 1 projects included measures such as lighting that do not require a change in a building's footprint or design. Tier 2 projects encompassed multiple systems through simple interventions or advanced measures that demand alterations to the building design, such as upgrading HVAC and installing advanced lighting controls. The design of the OBF program pushes the bounds of Rulebook definitions for population NMEC, and measures play a role in that.

4.2.5 M&V plans

The NMEC Rulebook requires that PAs establish M&V plans and calculation software before the start of population NMEC programs, and they apply these plans consistently once the programs begin. The Rulebook also recommends and encourages, though does not require, the use of savings approaches that are publicly vetted, such as CalTRACK methods, and calculation software that is public and open source. PAs that choose to use custom or proprietary software must submit these to the CPUC for pre-approval.

The PA outlined initial M&V plans for the OBF and the P4P programs in its Advice Letters (3697-G/4813-E and 3697-G /4812-E, for OBF and P4P, respectively). It recommended that the residential P4P M&V plan serve as a general framework, with detailed M&V plans to follow based on the customer types that enroll in P4P programs and the measures the programs install. The general framework proposed the use of a pre/post analysis based on a quasi-experimental design. The PA's OBF M&V plan provided in the Advice Letter also outlined a similar approach to estimate OBF program savings. Since these initial plans, the PA has issued updated M&V plans for the OBF and P4P population NMEC programs under evaluation in this report. The PA released the first updated version of the OBF M&V plan in July 2021.[71] It released the second revision version in August 2023.[72] The PA also provided an updated P4P M&V plan in May 2023.[73]

As part of our evaluation of the Rulebook's M&V requirements, we used the descriptions of the M&V plans provided in the Advice Letters, the PIPs, the PA Early M&V report for PY2019 P4P programs,[74] the M&V code set, and information from PA program staff interviews. To the extent possible, we also reviewed the information provided in the latest OBF and P4P M&V plans, released in May and August 2023.[75] While the programs fulfill the requirement of the Rulebook that population NMEC programs have M&V plans in place, due to the substantive change in M&V approaches the PA used to estimate the energy

[71] PG&E. "On-Bill Financing Alternative Pathway Program-Level M&V Plan." cedars.sound-data.com, Revised 07/16/2021.https://cedars.sound-data.com/documents/download/2241/main/.

[72]PG&E. "On-Bill Financing Alternative Pathway Program-Level M&V Plan." cedars.sound-data.com, Revised 08/08/2023. https://cedars.sound-data.com/documents/download/2882/main/.

[73] Pacific Gas and Electric Company. "Residential Pay-for-Performance Program. Measurement & Verification Plan." cedars.sound-data.com, "Residential Pay-for-Performance Program." https://cedars.sound-data.com/documents/download/2852/main/.

[74] Demand Side Analytics. "Early M&V Report for Program Year 2020 Residential Pay-for-Performance Program." pda.energydataweb.com, July 12, 2022. Early M&V Report for Program Year 2020 Residential Pay-for-Performance Program (energydataweb.com.

[75] The 2023 updates are substantially different and bear little resemblance to the early plans.

efficiency impacts, they do not meet the requirement that M&V plans get set before the start of such programs and applied consistently once the programs are underway. However, these changes reflect program designs that predate the NMEC Rulebook.

Given the changes in the early M&V plans that the PA proposed, including the complete revision of the PA's M&V code set that reflects considerably different approaches, we based our assessment of the M&V plans on validated savings, code review, and PA interview responses. Using these tools, we assessed the steps the PA took to establish robust M&V plans, barriers to applying population NMEC requirements, program influence and attribution, and cost-effectiveness. The following information applies to all four programs unless noted otherwise.

4.2.5.1 Establishing robust M&V plans

Since NMEC programs are in nascent stages, the PA faced various challenges to establish a robust M&V process. We present an overview of the various challenges the PA faced and the approaches it adopted to address them.

Tracking NMEC feasibility and alternatives in case of failure. For P4P programs, implementers employed a customer screening process to establish NMEC feasibility that the PA subsequently reviewed. Since residential projects in these programs were typically small in scale, the PA did not conduct extensive upfront screening of such participants but verified if they had sufficient data and did not acquire EVs or on-site solar during the baseline period. The PA M&V process also excluded P4P sites if they had poor baseline models whose CVRMSE was above a pre-determined threshold (above 1 for the programs under evaluation). In cases where the NMEC methodology was not feasible or unsuccessful for a site, the M&V process assigned savings to the site based on a similar cohort group's savings estimate.

For OBF programs, the PA implemented a screening process for loans above $25,000, which the program applied retroactively to projects from 2019 onwards. This process involved evaluating data sufficiency, changes in solar status, CVRMSE checks, and meter size to ensure that NMEC methods were viable. If the energy consumption of a meter represented more than 4% of the energy consumption of the population, the OBF M&V process filtered it out since it could have an outsized influence on the savings.[76] However, it was challenging to determine the population size at the outset of the year, which made filtering out such sites impossible until the performance period. The PA M&V process also excluded sites whose solar status changed during the baseline period and had poor models during this period. It assigned savings to these sites by applying the realization rate of non-excluded sites to their engineering savings estimates.

Approach to forecast savings. The P4P programs relied on historical savings estimates and deemed values to forecast savings. HomeIntel, which involved behavioral interventions, used estimates from previous program years to predict savings. The P4P programs that involved retrofits relied on engineering estimates and white papers to forecast savings.

For OBF, the trade professional/contractor responsible for each project provided an engineering estimate of the savings, which an Investor Confidence Project (ICP)-credentialed quality assurance (QA) reviewer validated before and after installation. OBF used this validated savings estimate to forecast program savings.

Approach to determine peak impact and hourly savings shapes. The PA has incorporated the DEER method in its M&V plan and uses it to determine peak impact. The PA offers program implementers an additional payment (kicker) for peak savings. Although it explored blended/weighted load shapes to establish hourly savings shapes, the process became complex, and the PA decided not to pursue this approach. The PA is exploring customized load shapes for the OBF program since it is larger than the P4P programs and requires a long-term solution.

[76] Gas programs were relatively uncommon for OBF, as the program primarily focused on lighting programs. Additionally, gas meters tend to be larger, making it easy for a single project to represent more than 4% of the population's energy consumption.

Method to determine effective useful life (EUL). The PA uses an EUL of 1 year for P4P behavioral programs. For P4P programs that involve retrofits, the PA uses a weighted EUL based on the installed measure mix. Lighting is the predominant measure for the OBF program. The PA has a tool that uses DEER hours for lighting products to determine the average EUL for these measures. The PA requires developers to submit specification sheets for non-lighting installations, which engineers review to validate EUL.

Method to distinguish savings from random variation for projects with less than 10% savings. The 10% rule of thumb pertains to site-level NMEC programs. The PA M&V process developed a tool that evaluated site savings and savings rates to ensure programs stay within a safe range of FSU for electric programs. One of the challenges with gas savings is that the data is not granular enough to provide an accurate savings rate, especially for non-heating days.

4.2.5.2 Barriers to applying population NMEC Requirements

The PA reported that calculating and meeting FSU requirements has been one of the biggest challenges. FSU calculations for gas and peak savings have been particularly problematic. The PA had difficulties in this area for both P4P and OBF programs.

Additionally, COVID-19 posed significant challenges to NMEC methodologies. The original CalTRACK method did not include comparison groups as part of the process. With the onset of the pandemic, the PA needed to control for the effect of this exogenous factor on energy consumption changes and program effects. The PA commissioned a study to explore comparison groups that could account for the impact of COVID-19 and other external factors, which did not require the transfer of large numbers of non-participant interval data to third-party M&V providers. The process resulted in granular profiles as controls, which meet the need to control for exogenous effects without transferring large volumes of non-participant data.

The PA also reported some minor data-related issues. It encountered some problems when developing the data collection system for OBF due to the complex metering structures of some participants. Additionally, a small sample of P4P participating homes appeared to be outliers because they were electrified and not correctly flagged. The PA was able to rectify both issues.

4.2.5.3 Program influence and attribution

The NMEC Rulebook provides guidelines for assessing program influence and attribution. The guidelines indicate that population NMEC programs can collect data to determine program influence or use CPUC-approved default NTG values. Although the population NMEC programs preceded the Rulebook, their design elements incorporate approaches (use of default NTG values) that comply with it.

In our interviews, the PA program staff noted that the P4P programs implemented comparison groups post-COVID-19 in 2020. They indicated that while comparison groups cannot eliminate free-ridership, they can mitigate its effects. They also noted that the PA applies a default NTG value of 85% for residential P4P NMEC programs to adjust for program influence. The updated PA M&V plan for P4P programs indicates that the PA will continue using a default NTG value of 85% to account for program influence.

During the interview, the PA staff also stated that OBF programs use comparison groups to address the effect of free-ridership. The PA also uses the default non-residential NTG value of 95% for OBF programs, approved in CPUC Resolution E-4952, for the same purpose. The PA's updated OBF M&V plan indicates the continued use of this default NTG for OBF program attribution.[77]

[77] Please note that the PA M&V results reflect savings estimates with and without the default NTG values for the OBF program.

While the programs comply with the Rulebook requirement for program attribution, DNV's estimated NTG values are generally below the default values that the programs use. Future population NMEC programs should undertake additional NTG research to understand the influence of these programs on participation.

4.2.5.4 Cost-effectiveness

The NMEC Rulebook stipulates that population NMEC M&V plans incorporate plans for cost-effectiveness. The PA program staff indicated that cost-effectiveness for the P4P programs has not been a top priority since the programs are relatively small in scale. Cost-effectiveness is also not a primary consideration in the screening process for these programs. However, in compliance with the NMEC Rulebook, the PA reports on cost-effectiveness at the portfolio level.

Similarly, the PA program staff noted the absence of cost-effectiveness screening for the OBF program. However, for projects exceeding $250,000, the PA calculates total resource cost (TRC) and total system benefit (TSB) scores during the pre-screening process. The scores are among the factors the PA considers when deciding projects for inclusion in the population NMEC program. The program staff also noted that OBF does not have a cost-effectiveness requirement because it is a market support program.

4.2.6 Program design criteria

The NMEC Rulebook specifies that "Population-level NMEC program designs must meet or exceed 90% confidence / 25% range Fractional Savings Uncertainty (FSU) as calculated using the ASHRAE methods at the daily level or using other methods that achieve at least the same levels of certainty." (NMEC Rulebook 2.0, II.2.C.1). The NMEC Rulebook is not sufficiently clear that this criterion refers to aggregate program level FSU in keeping with standard billing analysis precision measurements. The reference to ASHRAE methods, which are explicitly site-level calculations, could be misinterpreted as site-level eligibility rather than aggregate requirements. An aggregate FSU is more consistent with ensuring NREs are addressed appropriately in the population NMEC context. On the other hand, the program implementation plans did not provide how they predicted FSU.

4.2.7 Effective program management

DNV's evaluation team examined the methodology for tracking information, collecting data, and determining program performance deployed by the four NMEC programs to assess their adherence to the guidelines established in the NMEC Rulebook. Although the NMEC Rulebook mandates that P4P programs track relevant program information, it does not provide a specific roadmap, leaving it to the discretion of program administrators to design their program tracking systems. Interviews with the program administrators and implementors enabled the DNV evaluation team to collect relevant information about program tracking. We also leaned on the PIPs and Advice Letters to understand the NMEC requirements around this issue.

P4P programs aim to collect complete customer information and savings for effective program management. The PA has streamlined the data collection process. Relevant program information such as claimed savings, program and implementation costs, and participant and contractor details are all tracked in the PA's Energy Insights system and maintained in Salesforce. Implementers submit their monthly enrollments to Energy Insights, which is then transferred to the PA vendor to track primary activities and savings. The PA uploads the data and savings to CEDARS. The PA does not conduct claimable savings analysis on an ongoing basis but solely on the first-year savings, whereas it performs payable analysis on cumulative savings over two years.

For all four evaluated programs, the PA follows CPUC requirements to true-up initial claims by conducting a savings analysis at the end of each year. Meeting both internal and external deadlines and maintaining quality control can be difficult, given that calculating savings numbers can take weeks or even months. The PA staff noted that they often must rush to meet these deadlines.

Like P4P programs, OBF customer data is reported in Energy Insights and maintained in Salesforce. Lighting projects make up a significant portion of the projects in OBF. The PA has developed a tool to standardize the process for these projects. For non-lighting projects, OBF requires project developers or contractors to have engineering staff with ICP certification training, ensuring they understand the requirements for M&V. Contractors may also use third-party tools and methodologies, such as eQuest, to validate project results against actual meter-based savings.

All four population NMEC programs can improve the tracking of information related to ongoing program eligibility, which has the potential to change during the performance period. These data, leading to some exclusions when calculating program savings, are essential to final savings estimates and need to be fully documented throughout the process. This ongoing tracking of customers is an added responsibility for implementers of NMEC projects above and beyond traditional energy efficiency programs. Improving the tracking and documentation of information that feeds into eligibility exclusions should be the next focus as the PA continues to develop its program management systems.

4.3 Assessment of gross methods and results

We provide our assessment of the P4P and OBF gross savings methods and results in this section. We examine if the methods used are sound, meet industry standards, and are replicable and transparent. We also assess whether elements of the analysis meet accepted criteria. These include sufficient documentation of customer attrition, sound data cleaning and outlier treatment, program savings that use at least 12 months of pre- and post-installation data, and savings that reflect current TMY weather data. Finally, we discuss the reasonableness of the evaluated savings and if these savings reflect expected levels of uncertainty and align with PA reported and final claimed savings.

4.3.1 Replicability and transparency of methods

Since validation of population NMEC programs with embedded M&V processes require replication of results, challenges that make such replication impede a successful evaluation. This section provides the challenges we faced in this area.

The current evaluation is the tail-end of a long process, which DNV embarked on to replicate and validate the PA M&V analysis and results for PY2019 to PY2021 population NMEC programs. The effort involved three separate replication endeavors.

The first involved replicating the PA's early P4P M&V report results for PY2019 P4P programs. DNV used the PA's data and code to generate and validate these results. This attempt was only partially successful because we did not receive data for a subset of the participants, which we uncovered following several rounds of communication with the PA.

The second replication involved the PA's PY2020 and PY2021 P4P program results. DNV only partially succeeded in replicating the site-level and aggregated results because we appeared to be using dated analysis data. Comparisons of PA and DNV site-level results indicated possible differences in granular profile data for two cohorts that were the source of most discrepancies.

The PA updated the analysis data following quality checks, which uncovered duplication of energy consumption for about a third of the participants included in the analysis, which required a third effort to replicate new results. These results also reflected substantial changes in the M&V modeling approach. DNV based its evaluation results on the updated data, modeling method, and code. DNV's third replication of these results was relatively successful and is the basis of the evaluation described in this report.

As the above demonstrates, the latest replications were closer than earlier attempts, indicating the evaluation process and the questions it raised helped tighten up and improve the M&V process. For example, the PA-vetted evaluation data thoroughly streamlined the M&V methods, and the PA M&V process paid closer attention to elements of the analysis to ensure they reflected generally accepted criteria. Such improvements included a better identification of movers and explicit checks for data sufficiency.

For the non-residential OBF programs, DNV received materials the PA used to generate M&V savings for OBF projects with claimed savings in PY2020. Among the materials we received to evaluate these claims was a methods document detailing the steps used to generate M&V savings, which included data preparation code snippets and a reference to an open-source code used to produce key model results. DNV also received a roster of participants with details about the intervention, including installation dates. Additional files DNV received included raw input and results data files to facilitate the replication of savings.

While the instructions were clear, replicating the steps to produce the intermediate steps required considerable time and effort and had the potential to compromise replicability. The biggest challenge in replicating results arose due to the unclear procedures the PA used to generate savings.

While the challenges of validating the PY2020 PA M&V values involved the absence of code for critical elements of the data analysis and the lack of clarity on the steps used to generate the savings provided in two results files, DNV received the code set and data to facilitate the replication of the PY2021 OBF results.

4.3.2 Assessment of PA M&V approaches

Methodological soundness. While the models used for the P4P and OBF programs meet the limited guidance in the Rulebook and conform with industry standards, the extreme challenges of replicating and validating the savings estimates limited the extent to which this evaluation could review these methods more comprehensively. However, it is essential to understand the processes applied by the PA M&V approach, including some novel aspects of the methodologies employed to determine both P4P and OBF savings.

The synthetic control (SC) approach is one of the novel approaches of the PA M&V process. The method is well-established in the econometrics literature but is new to energy efficiency evaluations. It offers an interesting addition to the evaluation toolkit. However, there are essential differences between standard applications of the SC approach and the application for NMEC. The basic SC method used in other applications produces a "synthetic" control by creating, what is effectively, a weighted average across multiple possible subjects that are similar to the participant or treated entity in every way but the treatment. The classic paper compared aggregate state-level statistics, for example, for both the treated state (CA with a cigarette tax) and the remaining states (without cigarette taxes).[78] The data representing the treated entity and the controls were at a similar level of aggregation and parameter estimates functioned as weights, creating a single "synthetic" control out of the multiple controls included.

The population NMEC M&V application of SC uses aggregate load profiles (granular profiles) created within segments defined by geography, size, solar status, and load shapes.[79] Granular profiles are embedded in a version of the time-of-week-and-temperature (TOWT) model. The model estimates participant baseline period load as a function of granular profiles, time of week indicators, and piecewise linear weather variables. As a result, we can classify this better as a hybrid approach. The aggregate controls cannot sufficiently explain variation in site-level loads as synthetic controls could in the original application. At the same time, the standard site-level weather TOWT model, typically used to capture all aspects of a site's load, is applied in combination with the granular profiles with the recognition that these profiles are weather-correlated. These hybrid SC models can predict performance-period load using performance-period granular profiles and actual

[78] Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." Journal of the American Statistical Association 105 (490): 493–505. https://www.tandfonline.com/doi/abs/10.1198/jasa.2009.ap08746

[79] Non-residential granular profiles are primarily based on industry segments.

weather. However, normalization to typical weather requires a second (again, novel) regression of performance-period savings estimates as a function of weather.

With the granular profiles, the SC approach incorporates a facsimile of a comparison group and produces an estimate of the performance period baseline load used to calculate savings. While each participant's savings estimate reflects a control, it is at the aggregate level where these population NMEC methods produce an unbiased estimate of savings with a certain level of precision that accounts for exogenous change and NREs. How fully or effectively the SC approach does that is yet to be fully explored. The potential shortcomings of the long-standing, matched comparison approach to addressing this challenge are well-explored and understood. For example, there is a recognition that different consumption trends between the self-selected participants and the chosen comparison group are possible and will lead to biased savings estimates. Nothing about the SC approach indicates it would be immune to this potential shortcoming. It is possible to the extent that matching has any potential to identify comparison customers whose consumption will move through time more like the participants, that the comparison group approach would tend to be less systematically biased. Comparing the implications of this challenge across methods quantitatively is not straightforward but should be pursued.

The accuracy of savings precisions estimated using the SC method also deserves scrutiny. A critical aspect of aggregate billing analysis methods is developing estimated precisions that fully reflect the underlying variability in the data. The aggregate savings estimates in a matched comparison group context fully entrain the errors of the matched comparator. Using granular profiles in the SC or a matched comparison group context may underrepresent the level of comparator variability and, as a result, overstate resulting precisions. The savings estimates may not appropriately account for the variability associated with the mean granular profiles. They may also not account for the lack of independence with the multiple use of granular profiles across participants. Precision levels should indicate the extent to which program participant counts have reached a level to address the variability inherent in participant and comparator loads. The Rulebook sets an expectation for the degree of NMEC results precision and the accuracy of their estimation.

The PA motivated the SC approach and the use of granular profiles necessary for the SC approach, to address concerns regarding implementer access to individual non-participant hourly consumption data. While the granular profiles offer one way to address these concerns, they require an additional and ongoing technical process to pull those random subsamples, process them correctly and keep them appropriately updated as the population changes over time. Initially, the PA faced challenges maintaining this process and had to provide updated granular profiles after issues were discovered. The granular profile process requires an additional technical process with its own associated complexity and risk in what is already a data-intensive process. Further analysis of the granular profiles being used for the SC method has already been identified as an area for further analysis.

The PY2021 OBF program used the SC approach discussed above. The PY2020 OBF program used a difference-in-difference (DID) approach that was somewhat more standard but used granular profiles rather than matched comparison sites. Since there could be substantial magnitude differences in energy consumption between participants and matched granular profiles, the method used a proportional approach to apply the DID.[80] The use of granular profiles in the DID context was motivated by data access restrictions created by the PA, which wanted to limit the transfer of large volumes of non-participant customer data for matched comparison analysis. While replacing a matched comparator with a granular profile solved this immediate problem, it moved the method outside any existing norm, raising issues parallel to those discussed above regarding synthetic controls. The proportional DID was necessary because of the use of GPs but was more likely to increase variation. None of the population NMEC programs used this method but could in the future, requiring its assessment in parallel with synthetic controls.

[80] The approach matched granular profiles to participants and used their percent pre-post consumption installation differences to adjust savings, making it possible to incorporate controls in similar operating conditions as participants but with vastly different consumption levels.

Customer attrition documentation. P4P did not provide a full accounting of participant data attrition. DNV's construction of the process revealed differences in the number of final participants used in the evaluation for PY2019 while those for PY2020 and PY2021 were the same. Without a complete accounting of attrition at each possible step, we were unable to account for the source of the discrepancies for PY2019, requiring a lengthy communication with the PA to reconcile the differences. In general, a complete list of all excluded sites, the stage of the analysis at which they were excluded, and the reason for exclusion would have facilitated the vetting of the results provided.

While the PY2020 OBF M&V process provided sufficient customer attrition documentation, DNV could not replicate the reported attrition since the provided data and analysis steps did not support the reported attrition. The OBF 2021 attrition seems to indicate a selective inclusion of sites for analysis, with an effect on the estimated savings and their level of statistical significance.

Analysis data preparation. The third PA M&V process included a P4P data cleaning process centered on checks for duplicates. The previous processes had included additional standard checks and routines required to prepare energy consumption data for analysis, including checks for missing data and zero reads. These processes also included outlier identification and removal. Because the additional data cleaning steps were not provided to DNV in the third M&V process, it is not clear if the P4P data preparation meets standard requirements.

The PY2020 OBF analysis also provided a robust data cleaning process, but DNV used the prepared analysis data and did not reproduce this data based on the provided data preparation steps. Similar to the P4P case, the PA PY2021 OBF data preparation cleaning process only centered on checks for duplicates. The code DNV received did not include any steps the PA may have taken to conduct additional standard data preparation checks.

Analysis data sufficiency. The PA's P4P analysis checked for the availability of at least 12-months of pre- and 12-months of post-intervention period data. Thus, the PA M&V process used P4P baseline models based on adequate pre-period data. Participants' energy savings also reflected the level of post-period energy consumption data they had.

Similarly, the PY2020 OBF baseline models and energy savings were based on complete and pre- and post-intervention period data. On the other hand, PY2021 OBF participants without adequate pre- and post-period data appear to have been assigned savings, which violates acceptable criteria for determining program savings.

Weather-normalization of savings. The P4P programs provided weather-normalized energy savings estimates that reflect TMY weather data. The PA used CZ2022 normal weather data provided on CALMAC to accomplish this. Similarly, the PY2021 OBF-modeled savings were provided on a TMY basis. However, the PY2020 OBF savings did not report weather normalized savings using TMY data as required by the NMEC Rulebook. The M&V methods and results document that DNV received indicated that the use of reporting period temperature results in weather-normalized counterfactual energy consumption and savings.

Reasonableness of estimated savings. The estimated savings for the P4P programs appeared reasonable, with weather-normalized savings within 10% of energy consumption. The PY2020 and PY2021 OBF average estimated electric savings were 2% and 7% of consumption, well below the preliminary estimated percent savings of 20% and above. As discussed in section 4.1.1, these estimates may be low due to NREs among big customers. It is also possible that preliminary estimates were unreasonably high. PY2020 OBF gas estimates of percent savings were similarly low at 3%, but PY2021 OBF gas savings represented 81% of consumption. Again, as discussed in the same section, the high PY2021 gas savings as a percent of consumption is an artifact of the small population and assigned savings. While multiple factors likely conspired to reduce performance-based savings from the preliminary engineering forecasts, it still seems likely that these estimates are unreasonably high.

Hourly savings shapes. The PA M&V results did not include estimates of hourly savings shapes. The PA staff indicated that they used DEER load shapes to determine hourly savings for P4P programs. The PA also indicated it will explore customized load shapes for the OBF program which involves larger projects. It will be important to use approaches that estimate hourly savings shapes for future population NMEC programs since total system benefits (TSB), which values energy saved at various times of the day differently, will be the metric used to evaluate the impact of EE programs.

Savings uncertainty. All but one PY2020 and PY2021 savings estimates for the HomeIntel and Comfortable Home Rebates programs meet the statistical precision requirements of (90/25) stated in the NMEC Rulebook. These represent an improvement over PY2019 estimates, which generally do not meet these requirements.

The HER program's PY2020 and PY2021 gas savings do not meet the NMEC Rulebook's precision requirements but have better than 90/50 precision, which is generally considered acceptable for a billing analysis. However, this program's electric savings are relatively low and imprecise. Population NMEC precision is calculated across customers and is sensitive to the number of customers included. The P4P PY2020 HER program has more than twice as many customers as the other P4P programs but does not meet the Rulebook's or generally accepted billing analysis precision requirements because of the relatively low savings estimates.

The OBF gas savings estimates are also highly imprecise. However, the OBF electric results are particularly problematic and contrary to the goal of obtaining reliable meter-based savings measurements.

4.3.3 Alignment of PA M&V savings with PA reported and claimed savings

During interviews with the PA's program staff, DNV learned that the PA's vendor tracks monthly enrollment and data. It uses these data to calculate savings to ensure timely payment for implementers. However, we did not receive any information on any ongoing monthly or quarterly savings results and supporting documents we requested. Although the PA's vendor calculates savings monthly, the PA does not have such reports and intends to build these, with monthly or quarterly cadence, starting in 2023.

As stated earlier, the revised PA P4P M&V results and the supporting data that DNV received did not match the PA's savings claims in CEDARS. As a result, DNV had to consider which savings numbers to use as the basis for evaluation – the CEDARS claims or the revised PA M&V. We elected to use the PA M&V results. These differing savings numbers complicated the process. A successful evaluation needs clear, accurate, and timely M&V savings that reflect those claimed with the CPUC.

For OBF, the PA trued up both PY2020 and PY2021 electric savings claims. Only the PY2020 PA M&V gross electric savings estimate matched the final claimed savings reported in CEDARS. The PA also provided PY2020 and PY2021 M&V gross gas savings estimates but did not report true-ups for gas claims in either year. The PY2020 and PY2021 PA M&V gas savings did not match the final CEDARS claimed gas savings.

4.4 Characterize participation

This section provides an overview of the participants whom P4P programs served from P2019 to PY2021, including their demographic profiles and experience with the programs. We also characterize customers served by PY2020 and PY2021 OBF programs through analyses of firmographics and participants' experience with the programs. We highlight the potential effects of these characteristics on program savings and program effects that participant experiences reveal.

We used survey data for the analyses. Appendix E provides details on how we expanded the sample to the population to develop the results presented in this section.

4.4.1 P4P program participant profile and experience

Table 4-20 provides a demographic profile of P4P program participants based on survey data.

Homeownership. A clear majority (80%) of program participants own the homes they live in. As noted in 4.2.1, all three P4P programs targeted single-family residential customers who tend to be homeowners in greater proportions than those in multifamily buildings. CHR also allowed the participation of those living in 2-4-unit buildings.

Dwelling size. Most participants also reported living in mid-sized homes, with 1000 to 2000 square feet. A higher proportion of HomeIntel participants live in bigger homes, above 2000 square feet than participants of the other programs. These homes are likely to use more energy and have more opportunities for energy savings.

Dwelling vintage. Most program participants also live in older homes, built before 1980. A higher proportion of HomeIntel participants live in older homes relative to participants in other programs, which affords additional opportunities for reducing energy consumption.

These findings reflect the high-opportunity homes targeted by HomeIntel and explain the relatively high electric and gas savings delivered by this P4P program.

Table 4-20. Residential profile of P4P program participants, PY2019 – PY2021

Own or rent	CHR (n=188)	HER (n=1,163)	HomeIntel (n=775)
Own	92%	84%	88%
Square footage	**CHR (n=176)**	**HER (n=1,115)**	**HomeIntel (n=759)**
2,001 or more	31%	29%	40%
1,001 to 2,000	64%	57%	60%
1,000 or less	5%	14%	0%
Home vintage	**CHR (n=179)**	**HER (n=1,117)**	**HomeIntel (n=758)**
Before 1980	55%	55%	64%
1980 to 1999	30%	26%	20%
2000 or later	16%	19%	15%

Data collected from P4P survey respondents also indicate that HomeIntel and HER served more affluent and educated customers. Table 4-21 provides the income and the levels of educational attainment of the three P4P programs. Higher proportions (over 50%) of HomeIntel and HER participants earned income over $100,000, while higher percentages of CHR participants reported income below $80,000.[81] More HomeIntel and HER participants also held college and advanced degrees than CHR customers (above 70% versus 53%). In contrast, CHR participants had lower levels of education than participants in the other two groups.

Of the three P4P programs in this study, HomeIntel achieved the highest savings. HomeIntel relies primarily on participant response to program messaging. The relatively higher proportion of affluent and educated customers targeted by HomeIntel can better navigate program information and achieve desired program outcomes. While the HER program involved participants in a similar demographic group as HomeIntel, it provided energy-saving kits (LEDs, showerheads, and faucet aerators) and technologies through an online marketplace, including home energy monitors and smart thermostats, that delivered the expected modest energy savings.

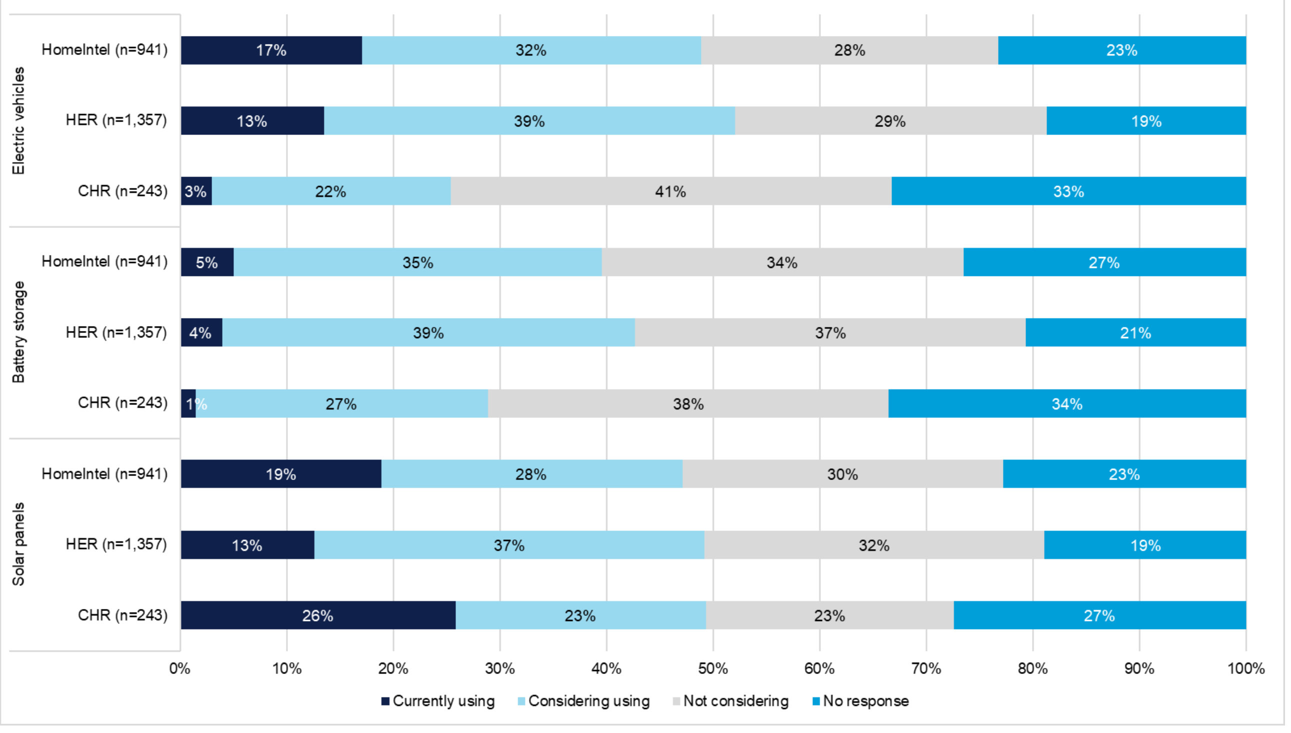
[81] The 2021 median income for California was $84,097. United States Census Bureau, “QuickFacts California,” census.gov, https://www.census.gov/quickfacts/fact/table/CA/PST045221

Table 4-21. Income and education levels of P4P program participants, PY2019 – PY2020

Income level	CHR	HER	HomeIntel
	(n=144)	(n=1,088)	(n=579)
$100,000 or more	24%	51%	57%
$80,000 to $99,000	14%	11%	11%
$79,000 to $40,000	36%	20%	19%
Less than $40,000	26%	19%	13%
Education level	**CHR**	**HER**	**HomeIntel**
	(n=173)	**(n=1,084)**	**(n=743)**
Bachelor or above	53%	72%	79%
Vocational/trade school or associate degree	24%	15%	13%
High school degree or less	21%	12%	7%
Other	2%	1%	1%

The survey also gauged P4P program participants' interest in clean technology. Figure 4-4 summarizes the proportion of P4P program participants who have adopted or are interested in three clean energy technologies.[82]

Figure 4-4. P4P participants' current and planned future adoption of other clean technologies, PY2019-PY2020

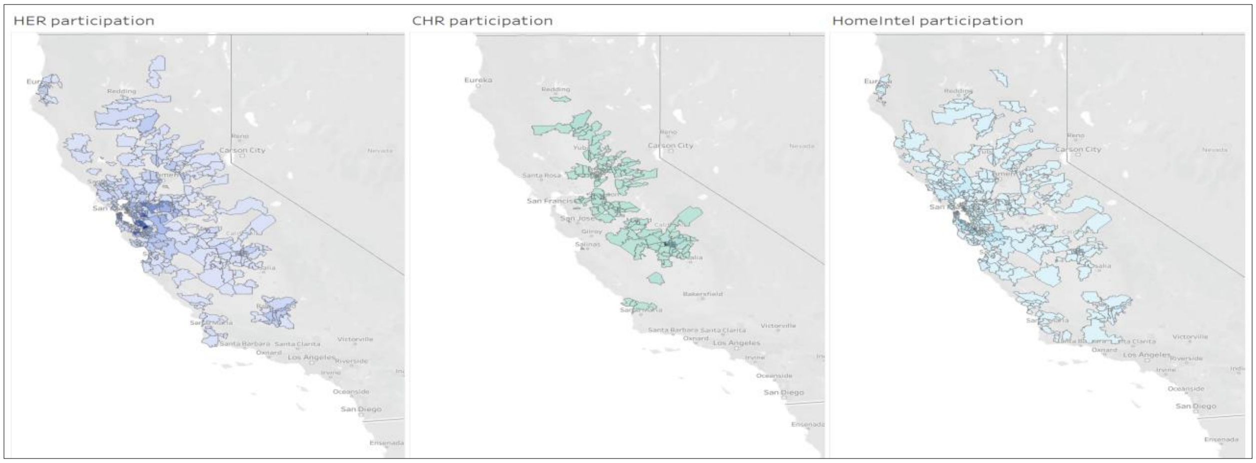


HomeIntel and HER customers have higher incomes and, as we would expect, higher percentages of EVs and battery storage than CHR customers. On the other hand, a higher percentage of CHR customers have more on-site solar than participants of the other two programs, which is unexpected given their relatively lower level of affluence than participants in the other two P4P programs. However, as indicated earlier, CHR targeted customers located in the inland climate zones 11, 12, and 13, where cooling needs are greater and could explain the higher prevalence of solar. The location of CHR

[82] The figure also includes the percent of non-responses in each category. While the percent of no response is relatively high, the pattern of clean technology adoption does not change when we only consider complete responses.

installations confirms this (Figure 4-5). By contrast, HomeIntel and HER participants are spread throughout PG&E's service territory, with a higher concentration in coastal climate zones 3 and 4.

Figure 4-5. Location of P4P program participants, PY2019 to PY2021



As noted above, HomeIntel and HER enrolled relatively affluent participants, and HomeIntel, in particular, targeted high-opportunity participants with high savings potential. CHR also targeted customers in hot climate zones where the direct install retrofit and HVAC maintenance it offered have the potential to deliver considerable levels of savings. The gross savings estimates are provided in section 4.1.1.1. Table 4-22 indicates that HomeIntel and CHR delivered notable energy savings, with average household electric savings of 5-8% and gas savings of 3-9%.

To understand the operation of HomeIntel better, we asked participants in the program about their experience with the online home audit and energy coach services offered by the program. Table 4-22 indicates that most participants received an online home audit and energy savings recommendations. In addition, about one-fifth of participants engaged with an energy coach to pursue targeted savings. Most of those who worked with a coach also received energy savings recommendations.

Table 4-22 Summary of HomeIntel audit and coach services participants received, PY2019-PY2021

Response	Online home audit	Online audit recommendations	Energy coach	Energy coach recommendations
Yes	94%	89%	16%	91%
No	6%	11%	74%	9%
Don't know			10%	

Approximately 76% of participants who verified receiving a home audit reported a range of recommendations the program provided them. Recommended actions included upgrading lighting to LEDs, unplugging equipment when not in use, adding insulation/weather stripping, power strips, timers, and smart plugs, installing solar panels, upgrading equipment including appliances, windows, and HVAC systems, and shifting energy use to off-peak hours. Table 4-23 shows the percentages of the top recommendations noted by participants.[83] Over 80% of those receiving online recommendations acted based on them. Given the percentage of HomeIntel participants who confirmed (89%), identified (76%), and followed the program's recommendations (91%), approximately half of HomeIntel participants undertook changes to affect energy consumption change.

[83] The total percent of recommended actions is above 100 because participants received more than one recommendation.

Table 4-23 HomeIntel program audit recommendations, PY2019 to PY2021

Program recommendation (n=561)	Reduce idle load	Upgrade lighting to LEDs	Shift energy use to off-peak hours	Upgrade equipment	Adjust Home/water temperature	Use power strips/smart plugs
Percent with recommendation	35%	19%	18%	16%	12%	8%

The participants who worked with an energy coach reported the same recommendations as those provided by the online audit, with additional tasks tailored to their circumstances. Examples include help identifying specific equipment using a lot of power, steps to follow when choosing upgrades, and recommendations to participate in another utility program to help with upgrade costs.

To gauge the effect of P4P programs, we asked participants about their experience with various aspects of the programs, starting with their satisfaction levels. We asked participants to indicate their level of satisfaction with different facets of the programs using a 5-point scale, where 1 is *Not at all satisfied,* and 5 is *Extremely satisfied*. As Table 4-24 shows, CHR and HER program participants are satisfied with the programs and the level of savings provided by them. About 80% of these participants reported satisfaction levels of 4 or 5. On the other hand, only 68% of HomeIntel participants reported the same level of satisfaction. They also reported low satisfaction with the program's energy savings, cost reduction, and non-energy benefits. Two sources of dissatisfaction emerged from verbatim responses provided by participants. Participants indicated that: (1) they were already implementing some of the program-recommended actions (2) some of these recommendations or installations were too expensive.

When asked about the benefits of the programs, a relatively high percentage of CHR customers (83%) reported improved comfort, safety, and convenience due to program installations, while only 15% of HomeIntel customers noted the same. Moreover, a higher percentage of CHR and HER customers reported no barriers to program participation compared to HomeIntel participants, who expressed experiencing cost barriers to implementing the suggested program changes.

Table 4-24. P4P program participants' experience, PY2019 to PY2021

Program satisfaction (rating of 4 or 5)	CHR (n=189)	HER (n=1,172)	HomeIntel (n=788)
Overall program experience	79%	80%	68%
Information and education provided by the program	71%	74%	70%
Experience with installation contractor/energy coach	83%		86%
Program equipment offerings	73%	75%	
Energy savings and cost reduction	74%	67%	55%
Non-energy impacts (e.g., increased comfort)	74%	60%	41%
Reported benefits	**CHR (n=197)**	**HER (n=1,221)**	**HomeIntel (n=1,159)**
Bill reductions/energy savings	40%	49%	37%
Improved comfort, safety, convenience	83%	35%	15%
No benefits experienced	12%	13%	0%
Don't know	10%	12%	9%
Barriers to participation	**CHR (n=191)**	**HER (n=1,118)**	**HomeIntel (n=784)**
No	83%	90%	72%
Yes	7%	10%	16%
Don't recall	10%	0%	13%

HomeIntel participants provided the program with low attribution and indicated that they would have undertaken program-recommended actions even in the absence of the program (section 4.1.2.1). The relatively low satisfaction levels and perceived benefits by participants of this program could be because they were already on the path to conserving energy with a downward trajectory in their energy consumption. It is possible they self-selected into the program and perceived that they would have achieved the measured level of savings on their own.

In addition, as indicated in Table 4-25, unlike CHR participants, HomeIntel participants reported environmental concerns as one of the primary reasons for their participation in the program. The table shows the most common reason for participation by all P4P programs is the desire to reduce energy and bills.

Table 4-25. Factors influencing P4P program participation, PY2019 to PY2021

Factors influencing participation	CHR (n=197)	HER (n=1,172)	HomeIntel (n=788)
Reduce energy use and bills	65%	84%	77%
Improve comfort, home safety, convenience	36%	33%	25%
Equipment failed or needed maintenance	52%		
Utility rebate/discount	30%	45%	
Reduce carbon emissions/ climate change/good for the environment	19%	47%	45%
Friend or colleague recommendation	2%	4%	2%

4.4.2 OBF program participant profile and experience

We collected information on the characteristics of the firms participating in OBF programs in this section.[84] We asked survey respondents when the organization in the current location was established, about the size of the organization participating in OBF, as measured by the number of employees, and the ownership status of the facilities where the organization operated.

Table 4-26 provides a summary of our findings. Over 40% of respondents indicated that the organization was established after 2000 and an additional 27% started operating in their current location between 1980 and 1999. As a result, the organizations participating in the OBF programs tended to be relatively new. Less than half (46%) of respondents reported that their organization employed 50 or fewer workers and tended to be small to medium-sized. Most (80%) of the respondents also indicated that the facility was owned, not rented.

Table 4-26. Characteristics of OBF program participants

Organization characteristics	Response
Year established (n = 68)	
2000 or later	43%
1980 - 1999	27%
1940 - 1979	20%
1910 - 1939	11%
Number of employees (n = 91)	
Fewer than 50	46%
51 to 250	27%
More than 251	3%
Don't know/No response	24%
Rental status (n = 91)	
Own	80%
Rent	17%
Manage	1%
Don't know	2%

[84] Appendix H provides a table summarizing additional OBF survey responses (by project type, loan size, and sector) that can help inform program design.

We asked survey respondents about the main activity of the facility where they installed the OBF-financed EE technologies. We used the information collected to characterize the market segment and organization type of customers participating in the OBF programs. Participants indicated the main activities of the facilities included agriculture, manufacturing, lodging, government administration, education and childcare, retail, grocery and convenience services, and offices and professional services. Based on this information, we classified the facilities into three main organizational types, including small and medium commercial, institutional (such as government facilities and schools), and large commercial, industrial, and agricultural. Table 4-27 summarizes the percentage of OBF participants by organization type. The majority (44%) of OBF participants were small and medium businesses.

Table 4-27. Percent of OBF Program participants by organization type

Organization type	n = 99
Small and medium commercial	44%
Institutional (government, schools, religious, etc.)	28%
Large commercial, industrial and agricultural	15%
No response	12%

We also asked OBF program participants what percent of their operating costs energy accounted for. As Table 4-28 indicates, most respondents (27%) with knowledge of the topic reported that energy accounted for 21% to 50% of their operating costs. It appears that energy constitutes a sizeable portion of the operating costs of most participants, which explains why the drive to reduce energy costs was a motivating factor for installing the OBF-financed EE measures.

Table 4-28. Percent of the cost of energy in operating costs of OBF participants

Cost of energy in operating costs	n=91
21% to 50%	27%
11% to 20%	22%
More than 50%	10%
Less than 10%	6%
No response	5%
Don't know	31%

DNV collected information to understand the program experience of participants in the PY2020 and PY2021 OBF programs. We used the information to understand the channels through which participants learned about the program, the reasons why they completed the EE program, chose OBF to finance the program, and their satisfaction with various facets of the programs.

Table 4-29 indicates that almost half (45%) of survey respondents indicated learning about the OBF program from PG&E sources (PG&E account manager, bill insert, or website). Project developers or contractors are also another important source of information on OBF, with a third of participants indicating having learned about the program through these actors.

Table 4-29. Primary source of information about the OBF Program

OBF information source	n = 94
PG&E sources - account manager, bill insert, or website	45%
Contractor, installer, or project developer	31%
Don't know	21%
From a friend or colleague	3%
Own research	1%

As Table 4-30 indicates, most respondents (84%) cited a desire to reduce energy usage as the primary motivation for completing the energy efficiency project. The second most common, but far less prevalent, reason for completing the project was the desire to reduce the organization's carbon footprint. Reasons related to equipment replacement or upgrades were far less common.

Table 4-30. Primary reason for participating in the OBF Program

Reason for project	n = 89
Wanted to reduce my organization's energy usage	84%
Wanted my organization to be more environmentally friendly	12%
The equipment broke and there was an immediate need for repair or replacement	1%
Upgraded the equipment to accommodate my organization's expansion	1%
Don't know	1%
No response	2%

DNV also asked about the primary reason for using OBF to finance the project. Table 4-31 summarizes the responses. A significant majority of the survey participants who answered the question (34%) cited the absence of upfront costs as one of the primary reasons for pursuing OBF. In addition, 19% and 16% mentioned energy savings and convenience, respectively, as their motivation for choosing the loan option. Furthermore, 9% of the respondents indicated they chose OBF because of a contractor recommendation.

Table 4-31. Primary reasons for using OBF to finance the EE program

Reasons for OBF	n = 32
No upfront Costs	34%
Energy savings/Bill neutrality	19%
Convenience	16%
Contractor suggested	9%
Do not remember	9%
Covered the full cost of the project	6%
Covered Lightning projects	3%
Other	3%

Overall, program satisfaction was consistently moderate. In response to questions about how satisfied they were with various program aspects (on a scale of 0 to 10, with 0 being "not at all satisfied", and 10 being "extremely satisfied), respondents provided an average rating of 7 to each. DNV classified responses to program satisfaction questions into three, with scores of 0 to 4 categorized as "Not satisfied", 5 and 6 as "Neutral," and 7 and above as "Satisfied or Very Satisfied." As Table 4-32 indicates, 71% to 74% of respondents are satisfied with different program features, relatively small proportions of 2% to 3% are neutral, additional 2% to 3% are not satisfied. Approximately 20% of respondents did not indicate their level of satisfaction.

Table 4-32. Satisfaction with different facets of the OBF project

Level of satisfaction	Ease of completing application	Amount of documentation for application	Time for application approval	Interaction with PG&E	Interaction with the contractor	Time to complete project	Charges on the bill
Satisfied or very satisfied	74%	71%	72%	74%	73%	72%	72%
Neutral	2%	3%	2%	3%	2%	2%	5%
Not satisfied	3%	2%	3%	2%	2%	2%	2%
Don't know	9%	9%	9%	9%	11%	11%	9%
No response	12%	15%	14%	13%	12%	12%	14%

5 CONCLUSIONS AND RECOMMENDATIONS

Despite the relative newness of population NMEC programs, impact results provide evidence of the potential of the approach. All population NMEC programs faced challenges with at least some aspects of the embedded evaluation methodologies and translating those results into saving claims. However, this evaluation demonstrates that the performance-based programs delivered by the PA were consistent with most NMEC Rulebook expectations and the savings validated for the residential P4P programs were generally better than savings from similar programs implemented recently in California. Furthermore, while there were various challenges, the overarching goal of developing empirically-based savings estimates that minimize ratepayer risk was met. In total, these results represent a successful, if tentative, step to demonstrating the substantial potential of NMEC programs for California.

Clarify necessary steps to take population NMEC to the next level. This evaluation considers the first programs developed under population NMEC (or ported over from HOPPS) and looks at a period impacted by the COVID pandemic, a systemic NRE unlike any previously seen. Concurrent with this process, a Working Group provided feedback on the existing NMEC Rulebook v.2.0, and revisions are underway. In the context of these developments, change, and external stresses, it is possible to focus on basic steps that will move NMEC to the next level.

Recommendation:

- Require up-to-date program implementation plans, program M&V plans, and final M&V reports prior to evaluation.
- Require timely savings claims in CEDARS consistent with internal M&V results.
- Require a package of internal M&V code and data documented to make evaluator replication straightforward.
- Offer more explicit guidance on eligibility requirements, for example, no addition of solar generation during the program period.

Explore and address possible risks in the NMEC process to ensure reliable and vibrant NMEC programs going forward. Many of the issues identified in this evaluation can be explained by the basic technical challenge of embedding the M&V function as part of the program implementation and the unprecedented challenge to both program implementation and evaluation caused by COVID-19. There remain areas of potential risk that could be problematic for all parties involved that deserve further and ongoing attention.

Recommendation:

- **Rules and the application of rules need to continually evolve to address challenges related to the precision of savings estimates and the potential for misuse of NMEC methods**.
  - Programs that use population NMEC methods should demonstrate that they can appropriately address the full range of NRE risks (such as changes in building occupancy) that could lead to potential over- or under-estimation of savings.
  - Rules that address customer population variability should be in place to address the possibility of large customer NREs that undermine a program's savings and precision. This may require redefining the FSU calculation for population NMEC. More generally, this includes developing methods for identifying strategic NREs if they occur and addressing them appropriately.
  - There should also be rules that address new onsite solar during the baseline or performance period for OBF.
- **The suitability of NMEC hourly savings for the application of avoided cost shapes for 2024 needs to be fully vetted.**

- While this evaluation focused on kWh and therm claims, in 2024, all electric claims will be based on hourly results. This shift requires a thorough examination of hourly savings methods and results to determine suitable approaches for the evaluation of total system benefits.
- In addition, there ought to be a focus on the appropriate precision level for hourly, including peak demand savings estimates. The precision level should account for the limited number of hours over which peak savings estimates can be made and the portfolio size required to achieve these.

Program effectiveness. In general, the P4P programs appear to have delivered notable savings, particularly in light of recent evaluations that indicate lower savings achieved by similar non-P4P programs. Part of their success seems to be due to the more effective targeting of participants likely to maximize savings.

Recommendation:

- Despite the evident success, customer feedback indicates room for improvement in targeting messages to what is present at a participant's home and what the customer is willing to invest.

6 APPENDIX

6.1 Appendix A: Gross and net lifecycle savings

Gross and net lifecycle savings are in the attached pdf.

6.2 Appendix B: Per unit (quantity) gross and net energy savings

Per unit (quantity) gross and net energy savings are in the attached pdf.

6.3 Appendix C: IESR−Recommendations resulting from the evaluation research

Study ID	Study Type	Study Title	CPUC Study Manager
Group A: CALMAC ID CPU0365.01	Impact Evaluation	Population-Based NMEC Programs - Program Years 2019 - 2021	Coby Rudolph

Rec #	Program or Database	Summary of Findings	Additional Supporting Information	Best Practice / Recommendations	Recipient	Affected document or domain
1	Population NMEC Programs	Clarify necessary steps to take population NMEC to the next level. This evaluation considers the first programs developed under population NMEC (or ported over from HOPPS) and looks at a timeframe impacted by the COVID pandemic, a systemic NRE unlike any previously seen. Concurrent with this process, a Working Group provided feedback on the existing NMEC Rulebook v.2.0, and revisions are underway. In the context of these developments, change, and external stresses, it is possible to focus on basic steps that will move NMEC to the next level.	Sections 4.1 and 4.2	• Require up-to-date program implementation plans, program M&V plans, and final M&V reports prior to evaluation. • Require timely savings claims in CEDARS consistent with internal M&V results. • Require a package of internal M&V code and data documented to make evaluator replication straightforward. • Offer more explicit guidance on eligibility requirements, for example, no addition of solar generation during the program period.	CPUC, All PAs	NMEC Rulebook, Program design consideration
2	Population NMEC Programs	Explore and address possible risks in the NMEC process to ensure reliable and vibrant NMEC programs going forward. Many of the issues identified in this evaluation can be explained by the basic technical challenge of embedding the M&V function as part of the program implementation and the unprecedented challenge to both program implementation and evaluation caused by COVID-19. There remain areas of potential risk that could be problematic for all parties involved that deserve further and ongoing attention.	Section 4.2	• Rules and the application of rules need to continually evolve to address challenges related to the precision of savings estimates and the potential for misuse of NMEC methods. – Programs that use population NMEC methods should demonstrate that they can appropriately address the full range of NRE risks (such as changes in building occupancy) that could lead to potential over- or under-estimation of savings. – Rules that address customer population variability should be in place to address the possibility of large customer NREs that undermine a program's savings and precision. This may require redefining the FSU calculation for population NMEC. More generally, this includes developing	CPUC, All PAs	NMEC Rulebook, Program design consideration

Rec #	Program or Database	Summary of Findings	Additional Supporting Information	Best Practice / Recommendations	Recipient	Affected document or domain
				methods for identifying strategic NREs if they occur and addressing them appropriately. – There should also be rules that address new onsite solar during the baseline or performance period for OBF.		
3	Population NMEC Programs	Explore and address possible risks in the NMEC process to ensure reliable and vibrant NMEC programs going forward. Many of the issues identified in this evaluation can be explained by the basic technical challenge of embedding the M&V function as part of the program implementation and the unprecedented challenge to both program implementation and evaluation caused by COVID-19. There remain areas of potential risk that could be problematic for all parties involved that deserve further and ongoing attention.	Sections 4.1 and 4.2	• The suitability of NMEC hourly savings for the application of avoided cost shapes for 2024 needs to be fully vetted. – While this evaluation focused on kWh and therm claims, in 2024, all electric claims will be based on hourly results. This shift requires a thorough examination of hourly savings methods and results to determine suitable approaches for the evaluation of total system benefits. – In addition, there ought to be a focus on the appropriate precision level for hourly, including peak demand savings estimates. The precision level should account for the limited number of hours over which peak savings estimates can be made and the portfolio size required to achieve these.	CPUC, All PAs	NMEC Rulebook, Program design consideration
4	Population NMEC Programs	Program effectiveness. In general, the P4P programs appear to have delivered notable savings, particularly in light of recent evaluations that indicate lower savings achieved by similar non-P4P programs. Part of their success seems to be due to the more effective targeting of participants likely to maximize savings.	Section 4.4	• Despite the evident success, customer feedback indicates room for improvement in targeting messages to what is present at a participant's home and what the customer is willing to invest.	All PAs	Program design consideration

6.4 Appendix D: NTGR survey scoring

For the Comfortable Home Rebates, Home Intel, and Home Energy programs, DNV used a standard NTGR approach that assesses three dimensions of free-ridership: timing, quantity, and efficiency. The programs induce savings if they accelerate the timing of measure installation, increase the number of measures installed, or raise the efficiency level of what was installed.

The timing and quantity dimensions are relevant to all measures. Efficiency is relevant for certain measures and not for others. For example, ducts are either sealed or not, so there is not a variable level of efficiency.

Using these metrics in combination allowed us to fully assess the amount of savings that could be attributed to measures that participants would have installed absent program support. We assigned each respondent a score for each free-ridership metric based on their survey responses and combined those scores into an overall free-ridership score using the algorithms in Equations 1 through 4.

Equation 1: Free-ridership Scoring Algorithm for AC replacement, Heating Replacement, Water Heater Replacement, Heat Pump HVAC, and Smart thermostats,

*Free-ridership= FR_timing * FR_quantity * FR_efficiency*

Equation 2: EE Kits and Home Intel participants

Free-ridership= FR_Liklihood

Equation 3: All other measures (Likelihood response “Very Unlikely”)

Free-ridership= FR_Liklihood

Equation 4: All other measures (Likelihood response not “Very Unlikely”)

*Free-ridership= FR_timing * FR_quantity*

Program attribution or NTGRs are simply the complement of free-ridership and estimated as: NTGR = 1- Free-ridership.

Measure and program-level NTGRs derived from participant surveys are weighted by savings claims to compute measure and program attribution estimates.

6.4.1 Comfortable Home Rebates

Table 6-1 provides the CHR program NTG scoring rubric by survey respondent type.

Table 6-1. CHR free-ridership elements and NTG scoring rubric by survey respondent type

Free-ridership dimension	Measures applicable	Question wording	Answer	Free-ridership score
Likelihood	**All**	Q2. The energy-efficient measures you received through the [program] program cost approximately [LOW-END PACKAGE COST]-[HIGH-END PACKAGE COST] to complete. Without the program, how likely would you have been to install any of these measures at your own expense? Would you say…?	Very Likely	FR_Liklihood = 1
			Somewhat likely	FR_Liklihood = 0.75
			A 50/50 chance	FR_Liklihood = 0.5
			Somewhat unlikely	FR_Liklihood = 0.25
			Very unlikely	FR_Liklihood = 0
			Don't know	FR_Liklihood = .
Timing	**All**	Q1A. Without the program, when would you have installed the product?	At the same time	FR_T = 1
			1 to 24 months later	go to Q1B
			More than 24 months later	FR_T = 0

Free-ridership dimension	Measures applicable	Question wording	Answer	Free-ridership score
			Never	FR_T = 0
			Don't know	FR_T = avg. of other respondents for that measure type
		Q1B. Please specify the number of months between 1 and 24:	Record #	FR_T = (24 - Q1B)/24
			Don't know	FR_T = avg. of other respondents for that measure type
Quantity	**All**	Q2A. Without the program, would you have installed the same, fewer, or more of the upgrades?	Same	FR_Q = 1
			Fewer	go to Q2B
			More	go to Q2B
			Don't know	FR_Q = average of other respondents for that measure type
		Q2B. How many would have been installed without the program?	Record #	FR_Q = 1 - ((original quantity - Q2b) / original quantity))
			Don't know	FR_Q = average of other respondents for that measure type
Efficiency	**AC System**	Q3. We would like to know what influence the PG&E program had, if any, on your decision to install the new high-efficiency AC system. Without the program, which of the following would you have done?	I would have purchased a STANDARD (minimum) efficiency AC system	FR_E = 0
			I would have purchased an INTERMEDIATE (lower than program requirements but above minimum) efficiency AC system	FR_E = .5
			I would have purchased a HIGH (same or higher than program requirements) efficiency AC system	FR_E = 1
			I would not have installed an AC system	FR_E = 0
			Don't know	FR_E = average of other respondents for that measure type
	Heating System Replacement	Q4. We would like to know what influence the PG&E program had, if any, on your decision to install the new high-efficiency heating system. Without the program, which of the following would you have done?	I would have purchased a STANDARD (minimum) efficiency heating system	FR_E = 0
			I would have purchased an INTERMEDIATE (lower than program requirements but above minimum) efficiency heating system	FR_E = .5

Free-ridership dimension	Measures applicable	Question wording	Answer	Free-ridership score
			I would have purchased a HIGH (same or higher than program requirements) efficiency heating system	FR_E = 1
			I would not have installed a heating system	FR_E = 0
			Don't know	FR_E = average of other respondents for that measure type
	Water Heater Replacement	Q5. We would like to know what influence the PG&E program had, if any, on your decision to install new high-efficiency water heating equipment. Without the program, which of the following would you have done?	I would have purchased a STANDARD (minimum) efficiency water heating system	FR_E = 0
			I would have purchased an INTERMEDIATE (lower than program requirements but above minimum) efficiency water heating system	FR_E = .5
			I would have purchased a HIGH (same or higher than program requirements) efficiency water heating system	FR_E = 1
			I would not have installed a water heating system	FR_E = 0;
			Don't know	FR_E = average of other respondents for that measure type

6.4.2 Home Energy Rewards

Table 6-2 provides the NTG scoring rubric for the HER P4P program.

Table 6-2. HER program free-ridership elements and NTG scoring rubric

Free-ridership dimension	Measures applicable	Question wording	Answer	Free-ridership score
Likelihood	**All**	Q2. The package of products you received as part of a kit/through the online marketplace of the program cost approximately [LOW END PACKAGE]-[HIGH END PACKAGE]. Without the program, how likely would you have been to purchase this performed at your own expense? Would you say…?	Very Likely	FR_Liklihood = 1
			Somewhat likely	FR_Liklihood = 0.75
			A 50/50 chance	FR_Liklihood = 0.5
			Somewhat unlikely	FR_Liklihood = 0.25
			Very unlikely	FR_Liklihood = 0
			Don't know	FR_Liklihood = .
Timing	**All**	Q1A. Without the program, when would you have installed the product?	At the same time	FR_T = 1
			1 to 24 months later	go to Q1B
			More than 24 months later	FR_T = 0
			Never	FR_T = 0

Free-ridership dimension	Measures applicable	Question wording	Answer	Free-ridership score
			Don't know	FR_T = avg. of other respondents for that measure type
		Q1B. Please specify the number of months between 1 and 24:	Record #	FR_T = (24 - Q1B)/24
			Don't know	FR_T = avg. of other respondents for that measure type
Quantity	**All**	Q2A. Without the program, would you have installed the same, fewer, or more of the upgrades?	Same	FR_Q = 1
			Fewer	go to Q2B
			More	go to Q2B
			Don't know	FR_Q = average of other respondents for that measure type
		Q2B. How many would have installed without the program?	Record #	FR_Q = 1 - ((original quantity - Q2b) / original quantity))
			Don't know	FR_Q = average of other respondents for that measure type
Efficiency	**Smart Thermostat**	Q3. Smart thermostats come in a variety of models. There are BASIC models that cost about $120-$150 (e.g., Nest E and Ecobee 3 lite) and UPGRADED models that offer additional sensing technology and cost about $210-$250 (e.g., Nest Learning 3rd Gen and Ecobee 4). There are also programmable and non-programmable thermostats that cost from $20-$100. If the program didn't offer a smart thermostat in 2021, which model would you have likely purchased?	Would have purchased the BASIC model smart thermostat	FR_E = 0
			Would have purchased a standard programmable or non-programmable thermostat (e.g., without smart capabilities	FR_E = .5
			Would have purchased the UPGRADED model smart thermostat	FR_E = 1
			Would NOT have purchased a thermostat at all	FR_E = 0
			Don't know	FR_E = average of other respondents for that measure type

6.4.3 HomeIntel

Table 6-3 provides the NTG scoring rubric for the HomeIntel P4P program.

Table 6-3. HomeIntel program NTG scoring rubric

Free-ridership dimension	Measures applicable	Question wording	Answer	Free-ridership score
Likelihood	**N/A**	Q2. Without the program, how likely is it that you would have taken the action(s)?	Very Likely	FR_Liklihood = 1
			Somewhat likely	FR_Liklihood = 0.75
			A 50/50 chance	FR_Liklihood = 0.5
			Somewhat unlikely	FR_Liklihood = 0.25
			Very unlikely	FR_Liklihood = 0
			Don't know	FR_Liklihood = .

6.5 Appendix E: Sample design and post-stratification

OBF sample design. This section provides an overview on the sample design approach that was used for the OBF program. The first step of the sample design was to define the population which was all OBF projects from program years 2019-2021. The sampling unit for the design was MMBtu savings at the customer and premise level, which was the sum of electric and gas savings converted into a consistent unit. Multiple claims for each participant were aggregated into a single sample point. The population was then stratified based on participant MMBtu savings. A total of four strata were defined. Table 6-4 presents the OBF stratified sample design including the strata cut points, number of accounts in the population, number of accounts in the sample, MMBtu savings, and inclusion probability for each stratum.

Table 6-4. OBF Sample design

Program	Stratum	Maximum	Accounts	Total MMBtu	Sample	Inclusion probability	Weights
OBF	1	697	597	136,271	30	0.1	19.8
	2	1,844	162	180,570	30	0.2	5.4
	3	8,851	64	229,822	30	0.5	2.1
	4	36,191	10	161,765	10	1.0	1.0

OBF post-stratification. Weights were developed for each measure type to align with the net to gross survey which asked participants attribution questions by measure. Due to the uniformity of savings within measure groups, a simple random sample was approach was used instead of using stratification. Table 6-5 presents the final post stratification including the number of participants in the population and sample, total savings, and weight.

Table 6-5. OBF Post-Stratification Results

Program	Measure	Stratum	Population	Sample	Weight
OBF	HVAC	1	43	10	4.3
	Lighting	2	754	80	9.43
	Refrigeration	3	37	6	6.17

P4P post-stratification. We developed weights at the measure level. Due to the uniformity of savings for each measure, we used a non-stratified expansion to calculate the weights. This means that within each measure type, all responses had the same weight. To estimate program-level metrics, including NTGR, net savings, and any other characteristics of interest, we applied measure-level weights to combine the estimated values. Using the developed weights, we estimated the population totals using mean per unit expansion. This approach involved calculating the mean value for a sample of customers and multiplying it by the population count. Table 6-6 and Table 6-7 provides the post-stratification details for the CHR and HER programs including the number of accounts in the population, sample, and corresponding weight.

Table 6-6. CHR post-stratification results

Measure Group	Stratum	Accounts	Sample	Weight
AC Replacement	1	175	83	2.1
Air Sealing	2	437	46	9.5
Attic Insulation	3	394	46	8.6
Coil Cleaning	4	517	10	51.7
Deep-Buried Ducts	5	213	15	14.2
Duct Replace/Improve	6	845	87	9.7
Fan Control	7	521	41	12.7
Heat Pump HVAC	8	41	6	6.8
Heating Replacement	9	534	56	9.5
Motor Replacement	10	158	7	22.6
RCA	11	1493	31	48.2
Right-Sized Returns	12	281	18	15.6

Measure Group	Stratum	Accounts	Sample	Weight
Smart thermostat	13	2193	146	15.0
Water Heater Replacement	14	54	4	13.5

Table 6-7. HER Post-stratification results

Measure group	Stratum	Accounts	Sample	Weight
Aerator	1	12	20	0.6
EE Kits	2	120	1182	0.1
HAN	3	94	24	3.9
LED	4	731	163	4.5
Showerhead	5	37	6	6.2
Smart thermostat	6	2217	353	6.3

6.6 Appendix F: Survey instruments and interview guides

6.6.1 Residential P4P survey instrument

The residential P4P survey instrument used in the evaluation is included as a pdf attachment.

6.6.2 Non-residential OBF survey instrument

The non-residential OBF survey instrument used in the evaluation is included as a pdf attachment.

6.6.3 PA interview guides

PA interview guides used in the evaluation are included as pdf attachments.

6.6.4 Implementer interview guides

Implementer interview guides used in the evaluation are included as pdf attachments.

6.7 Appendix G: Distribution of CVRMSE values for all program estimates

The figures in this section provide the distribution of model CVRMSE values for the evaluated population NMEC programs by program, year, and fuel type. They feature violin plots that combine box and density (frequency) plots, which make it possible to visualize both the summary of the distribution (such as the range and median) and the concentration (density) of the values. The figures display values without the top and bottom 1% CVRMSE values to facilitate visualization and understanding of the concentration and variability of most model CVRMSE values.

Figure 6-1. Distribution of P4P model CVRMSE values

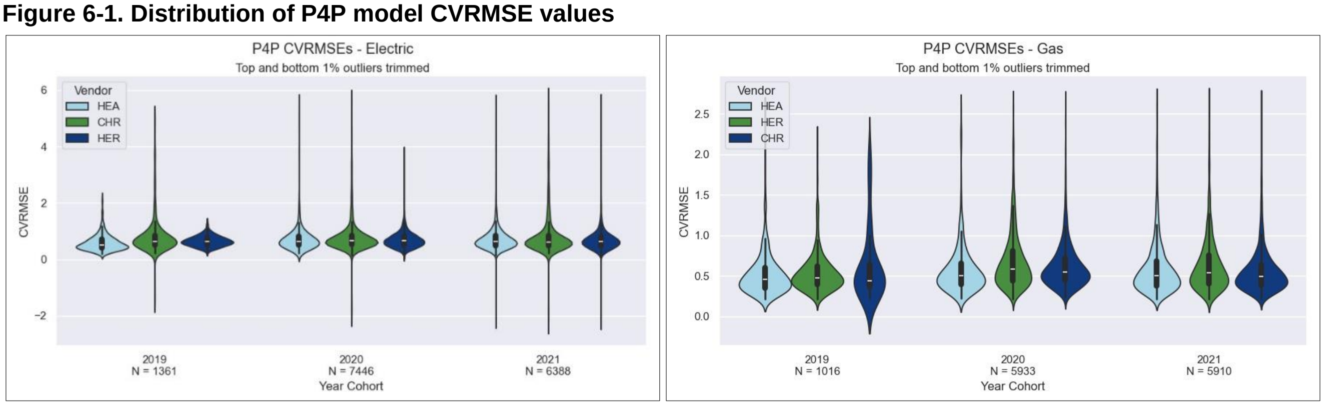
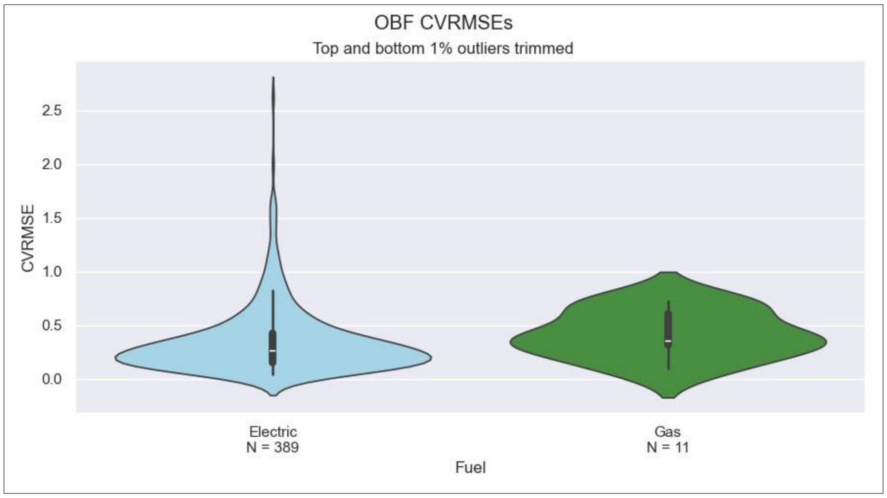


Figure 6-2. Distribution of OBF model CVRMSE values



6.8 Appendix H: Summary of OBF survey responses

The table below provides a summary of OBF survey responses that can help inform program design. We collected survey data at the project level since the program data we received was that level of granularity. The table provides survey response summaries by project type, loan size, and sector, which reflect project-level responses.[85]

Table 6-8. Summary of OBF survey responses by group[86]

Survey topic	Response	Project type			Loan size		Sector	
		Lighting (n = 78)	Non-lighting (n = 12)	Both (n = 9)	\$0 - \$50,000 (n = 42)	> \$50,000 (n = 57)	Institutional and SMB (n = 84)	Large C&I and Ag (n = 15)
OBF/rebate aware	Yes	100%	100%	100%	100%	100%	100%	100%
Aware of measure	Yes	100%	100%	100%	100%	100%	100%	100%
Used alternative financing	Yes	9%	22%	56%	14%	17%	15%	14%
	No	75%	66%	44%	76%	64%	70%	81%
	Don't know	15%	12%	0%	11%	19%	15%	5%
	Refused	0%	0%	0%	0%	0%	0%	0%
Alternative financing used [Select all that apply for respondents who said yes when asked if they used alternative financing]	Cash	46%	24%	33%	40%	36%	35%	67%
	Credit card	0%	24%	0%	0%	9%	0%	33%
	Utility rebates	15%	0%	0%	0%	18%	8%	0%
	Other financing	0%	0%	0%	0%	0%	0%	0%
	Other	54%	24%	67%	60%	45%	52%	67%
	Don't know	7%	76%	0%	0%	37%	17%	0%
	Refused	0%	0%	0%	0%	0%	0%	0%
Percent of project cost covered [Select all that apply]	OBF	99.7%	98.2%	100.0%	100.0%	98.2%	99.7%	98.5%
	Cash	0.0%	1.8%	0.0%	0.0%	0.7%	0.0%	1.5%
	Credit card	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	Utility rebates	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%
	Other financing	0.2%	0.0%	0.0%	0.0%	1.0%	0.2%	0.0%
Primary source of information	PG&E sources	47%	27%	43%	47%	41%	44%	52%
	Contractor	30%	51%	19%	34%	25%	34%	5%
	Friend or colleague	3%	5%	0%	0%	7%	2%	9%
	Own research	1%	0%	0%	0%	1%	1%	0%
	Don't know	20%	17%	38%	18%	25%	19%	34%

[85] Please note that for program attribution, we parsed out and collected survey data for individual measures from each project so that we could report NTG values at the measure level.

[86] We surveyed 94 projects out of a total of 816 program projects. The table provides the number of survey respondents by project type, loan size, and sector. The values in the table reflect responses from these survey respondents, but please note that not all survey respondents answered all the questions.

Survey topic	Response	Project type			Loan size		Sector	
		Lighting (n = 78)	Non-lighting (n = 12)	Both (n = 9)	$0 - $50,000 (n = 42)	> $50,000 (n = 57)	Institutional and SMB (n = 84)	Large C&I and Ag (n = 15)
	Refused	0%	0%	0%	0%	0%	0%	0%
Identified project before OBF	Yes	62%	84%	49%	63%	63%	67%	32%
	No	32%	5%	43%	31%	29%	27%	53%
	Don’t know	6%	5%	7%	6%	6%	5%	15%
	Refused	0%	5%	0%	0%	1%	1%	0%
Project cost approved before OBF [For respondents who said yes when asked if they identified projects before OBF]	Yes	79%	87%	100%	79%	84%	80%	100%
	No	11%	6%	0%	11%	7%	10%	0%
	Don’t know	11%	6%	0%	10%	9%	10%	0%
	Refused	0%	0%	0%	0%	0%	0%	0%
Aware of PG&E rebate	Yes	25%	31%	0%	20%	30%	24%	19%
	No	11%	12%	53%	15%	12%	13%	19%
	Don’t know	64%	57%	47%	65%	57%	62%	62%
	Refused	1%	0%	0%	0%	2%	1%	0%
Primary reason for completing project	Environmental concerns	12%	17%	0%	14%	8%	12%	9%
	Reduce energy use	85%	67%	100%	83%	86%	84%	81%
	Building code	0%	0%	0%	0%	0%	0%	0%
	Broken equipment	0%	11%	0%	0%	3%	1%	5%
	Accommodate expansion	1%	0%	0%	0%	2%	0%	5%
	Don't know	0%	5%	0%	0%	2%	1%	0%
	Refused	2%	0%	0%	3%	0%	2%	0%
Satisfaction with ease of completing OBF application	Very satisfied/satisfied	79%	66%	77%	76%	80%	76%	86%
	Neutral	3%	0%	0%	3%	1%	2%	0%
	Not satisfied	1%	22%	0%	0%	7%	2%	9%
	Don't know	9%	5%	0%	8%	8%	8%	5%
	NA/no response	9%	6%	23%	14%	3%	11%	0%
Satisfaction with amount of documentation for application	Very satisfied/satisfied	80%	66%	77%	76%	81%	76%	91%
	Neutral	3%	5%	0%	3%	3%	2%	5%
	Not satisfied	0%	17%	0%	0%	4%	2%	0%
	Don't know	9%	5%	0%	8%	8%	8%	5%
	NA/no response	9%	6%	23%	14%	3%	11%	0%
	Very satisfied/satisfied	77%	71%	77%	76%	77%	75%	91%

Survey topic	Response	Project type			Loan size		Sector	
		Lighting (n = 78)	Non-lighting (n = 12)	Both (n = 9)	\$0 - \$50,000 (n = 42)	> \$50,000 (n = 57)	Institutional and SMB (n = 84)	Large C&I and Ag (n = 15)
Satisfaction with time to get approval of application	Neutral	3%	0%	0%	3%	1%	2%	0%
	Not satisfied	1%	17%	0%	0%	6%	2%	5%
	Don't know	11%	5%	0%	8%	13%	10%	5%
	NA/no response	9%	6%	23%	14%	3%	11%	0%
Satisfaction with interaction with PG&E	Very satisfied/satisfied	80%	61%	77%	76%	80%	76%	91%
	Neutral	3%	5%	0%	3%	3%	3%	0%
	Not satisfied	0%	22%	0%	0%	6%	2%	5%
	Don't know	9%	5%	0%	8%	8%	8%	5%
	NA/no response	9%	6%	23%	14%	3%	11%	0%
Satisfaction with interaction with the contractor	Very satisfied/satisfied	79%	66%	77%	76%	80%	76%	86%
	Neutral	3%	0%	0%	3%	1%	2%	0%
	Not satisfied	1%	22%	0%	0%	7%	2%	9%
	Don't know	9%	5%	0%	8%	8%	8%	5%
	NA/no response	9%	6%	23%	14%	3%	11%	0%
Satisfaction with time to complete the project	Very satisfied/satisfied	77%	71%	77%	76%	77%	75%	91%
	Neutral	3%	0%	0%	3%	1%	2%	0%
	Not satisfied	1%	17%	0%	0%	6%	2%	5%
	Don't know	11%	5%	0%	8%	13%	10%	5%
	NA/no response	9%	6%	23%	14%	3%	11%	0%
Satisfaction with charges on the bill	Very satisfied/satisfied	80%	66%	59%	74%	81%	75%	91%
	Neutral	3%	5%	18%	5%	3%	4%	5%
	Not satisfied	0%	17%	0%	0%	4%	2%	0%
	Don't know	9%	5%	0%	8%	8%	8%	5%
	NA/no response	9%	6%	23%	14%	3%	11%	0%
Likely source of financing [Select all that apply]	Cash	94%	94%	66%	90%	97%	91%	100%
	Credit card paid in full	2%	0%	0%	0%	3%	0%	9%
	Interest bearing loans	3%	6%	34%	3%	9%	6%	5%
	Grant	31%	0%	0%	26%	25%	23%	43%
	None	5%	6%	0%	7%	2%	6%	0%
Year established	2000 or later	42%	41%	43%	49%	31%	45%	28%
	1980 - 1999	26%	39%	14%	21%	35%	24%	45%
	1940 - 1979	21%	20%	0%	20%	19%	21%	11%
	1910 - 1939	10%	0%	43%	9%	14%	10%	16%

Survey topic	Response	Project type			Loan size		Sector	
		Lighting (n = 78)	Non-lighting (n = 12)	Both (n = 9)	$0 - $50,000 (n = 42)	> $50,000 (n = 57)	Institutional and SMB (n = 84)	Large C&I and Ag (n = 15)
Cost of energy in operating costs	Less than 10%	7%	0%	0%	6%	6%	7%	0%
	11% to 20%	27%	0%	0%	26%	15%	23%	14%
	21% to 50%	19%	71%	46%	27%	27%	23%	53%
	More than 50%	10%	7%	7%	6%	15%	9%	15%
	Don't know	34%	23%	15%	29%	34%	34%	14%
	Refused	3%	0%	32%	6%	3%	5%	5%
Number of employees	Fewer than 25	30%	23%	53%	42%	13%	33%	19%
	26 to 50	16%	17%	0%	12%	19%	11%	39%
	51 to 100	11%	23%	0%	6%	21%	12%	9%
	101 to 250	19%	0%	0%	16%	15%	17%	5%
	More than 251	2%	11%	0%	0%	7%	2%	9%
	Don't know	19%	25%	47%	24%	17%	22%	15%
	Refused	3%	0%	0%	0%	7%	3%	5%
Ownership status	Own	82%	71%	77%	76%	88%	80%	81%
	Rent	15%	24%	23%	22%	9%	16%	19%
	Manage	1%	6%	0%	0%	3%	1%	0%
	Don't know	2%	0%	0%	3%	0%	2%	0%

6.9 Appendix I: Response to comments

We provide response to comments of the draft report we received in the table below.

Table 6-9. Response to comments

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
1	PG&E	General - Overall	PG&E appreciates the opportunity to review and provide comments and questions on this draft report. It is well-written and well-organized. PG&E looks forward to working with ED staff and its evaluation consultants on implementing the recommendations contained in this report to improve the implementation and administration of its programs that use the Population NMEC methodology.	Noted. Thank you.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
2	PG&E	General - NMEC Evaluation Process	Standard definitions for the terms "**program year**" and "**installation cohort**" for programs using the Population NMEC methodology are important, especially when installations of measures span more than a calendar year, and to determine the projects to include in the estimating first year savings. Would the evaluation team comment on these draft definitions as they would apply to programs using the Population NMEC methodology? **Program year:** the calendar year for which savings are calculated for the projects in a program that complete measure installation in that year. **Installation cohort:** The calendar month during which a set of projects will be analyzed as a group complete measure installation, because they have similar types of building characteristics (e.g., single-family homes with rooftop solar) equipment holdings, and drivers/levels of energy consumption.	We agree that standard definitions for "program year" and "installation cohort" for programs that use Population NMEC to make claims are important. The definition for program year is sound and we would modify the provided definition for installation cohort to reflect the similarity of intervention (program). The proposed modification is as follows: **Installation cohort:** The calendar month during which a set of projects in a **program** complete measure installation and will be analyzed as a group because they have similar types of building characteristics (e.g., single-family homes with rooftop solar), equipment holdings, and drivers/levels of energy consumption.
3	PG&E	General - NMEC Evaluation Process	Stakeholders could benefit if programs using the Population NMEC methodology used templates to document key operational processes, including project assessments, incentives, and savings calculations. Also, documenting the steps and timelines of "Early M&V" processes used to estimate and true up ex-ante savings could benefit stakeholders. Would the evaluation team offer suggestions in this regard, that might draw on their experience evaluating related distributed energy resource programs?	The Rulebook requires several documents for population NMEC programs, including an M&V Plan (with participant eligibility criteria and methods to calculate savings) and an incentive payment structure. Section 4.2 of the DNV report provides details on all the requirements. Rulebook revisions that provide updated reporting guidelines are now available. The revised guidance requires PAs to report initial claims by quarterly cohort and to true-up claims one year after installation based on M&V savings estimates. Please see https://docs.cpuc.ca.gov/PublishedDocs/Efile/G000/M520/K881/520881077.PDF DNV agrees templates that outline the required documentation and reporting timelines could be beneficial to improve consistency in population NMEC program reporting. It would be valuable for PG&E to draft such a template for ED review.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
4	PG&E	Section 1.4.2 - Incentive payment structure: OBF, Page 13	The draft report states "In the case of the OBF program, the PA provides a no-interest loan, and repayment is 100% based on performance-based savings. It is not clear whether implementers are paid based on performance, and because the OBF program pre-dates the NMEC Rulebook, it is not clear if this requirement would apply." PG&E would like to correct this factual error. The OBF Program *does not pay based on performance.* The OBF Program only pays the loan amount. No incentive or performance payments are made. Loan terms and monthly payment amounts are based on the Customer's pre-install estimated monthly energy savings from the retrofit project; the loans are fully repaid regardless of the intervention's measured performance.	We have made edits to correct this statement in the Executive Summary and the main report. The updates are as follows: Executive Summary update - In the case of the OBF program, the PA provides a no-interest loan to a participating customer based on estimated monthly energy savings, and the customer pays the loan amount regardless of performance. Thus, considering whether incentive payments reflect performance is not necessary. Main report update - The participating customer repays the no-interest OBF loan regardless of performance.
5	PG&E	Section 1.4.2 - Program design criteria, Page 14	PG&E agrees that the ambiguity in program design criteria continues to be a major source of confusion for both program staff and M&V vendors. Demand Side Analytics (DSA), PG&E's M&V consultant, has conducted simulations, using back-casted data from actual Res P4P and OBF participants and a bootstrapping procedure, to estimate FSU at the population level. PG&E/DSA believe that this empirical approach would yield useful information about the population FSU. Would the evaluation team recommend ways to address this ambiguity?	Our recommendation (provided in the report) is for the Rulebook to clarify that FSU requirements should be at the program and not at the individual project level for programs using population NMEC to claim savings. This standard is in billing analysis practice. Studies that provide insight into reasonable FSU for population NMEC programs could be valuable in determining the appropriate level of FSU for such programs. The revised Rulebook indicates the current requirement reflects the best available information and may be adjusted in the future as additional information becomes available.
6	PG&E	Section 3.2.2 - Program attribution approach, Page 28	PG&E recommends that future Population NMEC program impact evaluations further explore the interaction between the use of control adjustments and traditional net-to-gross ratio adjustments for free ridership, given that prior research indicates that control adjustments may account for *some* free ridership. For projects installed in PY 2019 and beyond, PG&E began including the use of control adjustments for the Population NMEC program in order to adjust gross savings estimates for systematic exogenous impacts on energy use (e.g., COVID-19) that impact savings estimates measured at the meter. There is considerable discussion within the industry on the extent to which comparison group adjustments of gross impacts (using the difference-in-differences or similar methodology) adjusts for free ridership (total, partial, and/or deferred free riders), even in the case of carefully-constructed comparison groups using customers who had participated in energy efficiency programs in prior years (see discussion in Chapter 8 of the Uniform Methods Project (Ken Agnew and Mimi Goldberg (2016). Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol. The Uniform	As discussed in the referenced UMP Chapter 8, while a comparison group may account for some net savings, it is a small fraction of the necessary adjustment. Particularly, a billing analysis estimate with a comparison group reflects no more than 5% of free ridership. Any free-ridership included in the billing analysis estimate is also likely dwarfed in magnitude by the self-selection bias potentially present for methods that use comparison groups or granular profile controls. Furthermore, billing analyses routinely do not remove savings between existing conditions and the current code baseline where appropriate, an adjustment that could only lower savings.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
			Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. Available electronically at SciTech Connect http:/www.osti.gov/scitech). There are factors that make it impossible for carefully-matched comparison to reflect fully the non-program changes among the participants. As a result, when comparison group change is netted out of the participant change, the netting will control for some—but not all—of the naturally-occurring measure implementation. This leaves some amount of free ridership in the final savings estimate, with the resulting estimate being somewhere in between net and gross savings. Given the foregoing, PG&E believes that, by discounting fully PG&E's estimates of gross savings that incorporate adjustment for exogenous trends of energy use by customers that are closely matched to program participants by the NTGR developed through the NTG surveys, PG&E's reported gross savings are being penalized to some extent. Would DNV please comment on the potential impact to the accuracy of the net savings estimate when using an adjustment to gross savings for free ridership using the result of the net-to-gross surveys in combination with using an adjustment to gross savings by using a comparison group to control for exogenous impacts of energy use?	
7	PG&E	Section 4.1.1.2 - OBF gross savings evaluation: Electric savings validation, Page 39	We observe that the draft report has a focus on statistical precision, and infer that this focus represents a policy priority. Given the focus on statistical precision, PG&E believes that it is important to have a shared understanding with respect to how Relative Precision (RP) and Fractional Savings Uncertainty (FSU) are computed and reported. Could the evaluation team provide more technical detail around the sentence "We calculate precision for population NMEC using the variation of the mean site-level savings across participants"? Specifically, would the evaluation team provide an equation for calculating precision (or at least a recommendation that the next iteration of the NMEC Rulebook provides equations for making these calculations)?	DNV calculated the FSU requirement using a formula for relative precision because uncertainty estimates are required at the program and not at the project level. The formula we used is specified as: $$\frac{t_{(1-\alpha),n-1}\sqrt{\frac{1}{n-1}\sum_i^n\left(\hat{S}_i-\bar{S}\right)^2}}{\bar{S}}$$ where: n = number of participants t = value from the t-distribution with $1-\alpha$ confidence level and $n-1$ degrees of freedom $\hat{S}_i$ = estimated savings for site i $\bar{S}$ = mean site-level savings
8	PG&E	Section 4.1.1.2 - OBF gross savings evaluation: Electric Savings Validation, Page 40	PG&E acknowledges that documentation about the removal of the two sites noted by the evaluation team could have been improved so as to not cause confusion. At the time of the analysis, a substantial non-routine event (NREs) (i.e., major operational changes) was discovered at these sites that resulted in a baseline consumption model that was not reflective of typical operations. At the time of	A feature of population NMEC is that the criteria identifying all such NREs are set in advance to avoid any perception of site-level savings (or lack thereof) driving program savings estimates. Defining categories based on EV additions or moving out is straightforward, but most NREs at non-residential sites cannot be as

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
			analysis, there was no process of assigned savings as there is now, documented in the M&V Plan. PG&E will ensure to update its documentation as well as ensure future documentation reflects program decisions made.	easily categorized. The possibilities of NREs are one of the reasons why claiming savings for non-residential interventions based on population NMEC methods is less than ideal. Based on the code review, the PA M&V process assigned savings to OBF projects with poor baseline models (CVRMSE > 1). If an NRE occurs during the baseline period, the baseline model's CVRMSE should identify it. If it does not, that raises the question of whether that benchmark serves our purpose and on what basis a neutral observer believes that NRE conditions are applied evenhandedly. DNV agrees that rules on participation eligibility and identification and treatment of sites with NREs, particularly for non-residential participants in population NMEC programs (such as those in OBF), should be laid out in M&V plans and followed in internal/PA M&V assessments to ensure no ad hoc removals, particularly those that have a material effect savings estimates and program-level precisions.
9	PG&E	Section 4.1.2.1 - Program attribution for P4P programs, Page 41	PG&E believes the current NTG survey battery, specifically the one related to self-reported likelihood of taking program-recommended actions(s) in the absence of the program (used as the sole basis for the NTG estimate for the HomeIntel Program) overestimates free ridership. The program theory of the HomeIntel program is to identify residential customers that have consistently used energy far in excess of what would be expected of typical customers occupying similar-sized homes within the same climate zone, gain a customer's cooperation to participate in the program, perform home audits and identify no-cost behaviors and low-cost measures that, theretofore, had not already been taken by the participant households. After adopting the recommendations of the program, the program staff continues to monitor the energy use at the residence, and if energy use does not demonstrate a substantial decline, then the program staff performs a second intervention to identify the root cause. PG&E expects that, if the majority of the program participants had known about the behavioral recommendations and low-cost measures prior to the participating in the home audits, that they would have implemented the changes prior to having participated (that is, during the baseline period or earlier). That being the case, we had expected a higher	While it is plausible that the self-reported likelihood of taking the program-recommended actions could be higher than is the case, participants indicated receiving recommendations for efforts they were already making as one of the two sources of dissatisfaction with the program. These reactions, provided independent of program attribution questions, corroborate the likelihood responses that were the basis of NTG estimates for the program. The Peters and MacRae study identifies a possible source of bias in self-reported program attribution, where respondents will overestimate the actions they would have taken without a program. However, the source of dissatisfaction reported by participants lends legitimacy to the HomeIntel program attribution estimates. Given these, it is unlikely that the evaluation significantly overestimated free ridership. Since default NTG estimates are not adjusted based on a single study, we recommend that future evaluation of this and related programs undertake further research to investigate the extent of free-ridership for such programs.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
			NTG for the behavioral measures implemented through the program. That certain biases are present in free ridership measurement has been acknowledged in the energy efficiency literature (for example, see the paper authored by Jane S. Peters and Marjorie McRae (2008) entitled, "Free-Ridership Measurement Is Out of Sync with Program Logic...or, We've Got the Structure Built, but What's Its Foundation" presented at the 2008 ACEEE Summer Study on Energy Efficiency in Buildings). One such bias is based on attribution theory, which suggests that "some people will be inclined to say they would have taken action in the absence of the program, even if some facet of the program was instrumental in the persuasion phases and resulted in the decision and action" and therefore would "overestimate free-riders" (Peters & McRae, 2008). In the case of a behavioral program, in which there is no financial incentive, it is possible that this bias exacerbated the issue. In light of HomeIntel's program theory and the potential bias in the NTG survey, would the evaluation team comment on the extent to which biases could have impacted responses to the survey question in a way that a possible overestimation of free ridership resulted for a behavior program? Given that the gross savings estimate has already been adjusted for partial free ridership using the comparison group methodology, would DNV comment on the extent to which gross savings has been adjusted for free ridership prior to applying the NTG ratio? In light of the foregoing, would the evaluation team consider adjusting the NTG for the HomeIntel program?	Moreover, program savings estimated using comparison groups do not provide estimates between net and gross savings. Please refer to the response to comment #8 for additional details.
10	PG&E	Section 4.1.2.2 - Program attribution for OBF programs, Page 43	Would the evaluation team provide PG&E with a full breakdown of survey responses that show how non-residential customers responded on each question, by sector, loan size, project type, and measure type? PG&E would find it valuable to see the specific responses to that it could apply any learnings from this breakdown to improve program design (e.g., project screening).	All individual survey responses are confidential, but DNV is able to provide aggregated responses by group that can help inform program design. Per the request, we have calculated weighted survey responses by sector, loan size, and project type where possible in Appendix 6.8 of the final evaluation report.
11	PG&E	Section 4.1.2.2 - Program attribution for OBF programs, Page 44	In Table 4-12, could the evaluation team provide an explanation as to how the weights per measure were determined, how the weights were applied to determine the NTG per measure, and then how the overall NTG was determined for the OBF-AP program?	For the OBF-AP program, we developed the weights by measure using a simple random sample methodology. In particular, the weight for each customer measure is the number of measures in the population divided by the number of measures in the sample. This process provided weights of 4.3, 9.4, and 6.2 for HVAC, lighting, and refrigeration. To expand the results of the sample to the population, we used a mean per unit (MPU) expansion. MPU calculates a weighted average net-to-

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
				gross ratio (NTGR) by measure and across all measures in the program, which is the sum of the weight times the NTGR divided by the sum of the weights.
12	PG&E	Section 4.1.2.2 - Program attribution for OBF programs, Page 46	PG&E recognizes the variability in the NTG values estimated for the OBF-AP Program. As noted in DNV's draft report (pg. 41), the 2018-2019 Cadmus study (CALMAC ID PGE0453.01) estimated an NTG of 0.94, ODC's PY 2018-2019 study found an NTG of 0.69 (see PDA for ODC's draft report https://pda.energydataweb.com/api/view/2780/CPUC%20OBF%20Evaluation%20Report_DRAFT_03092023_clean.pdf), and DNV estimated an NTG of 0.56. Could the evaluation team comment on the possible sources or explanations of the variability in program attribution for the OBF-AP program?	As we indicated in the report, DNV's evaluation relies on a more diverse and robust sample with 99 measure-specific responses (representing HVAC, lighting, and refrigeration) and a relative precision of ±9%. In contrast, the previous evaluations used limited samples of 23 (the 2018-2019 PG&E study) and 19 (the PY2018-2019 ODC evaluation). In addition, in the PG&E 2018-2019 study, the NTGR reflects a program attribution value of 100% from a single project, which accounted for over 80% of the program savings in the survey sample. Without this project, the average NTGR of the remaining projects included in the PG&E 2018-2019 would have been more in line with other recent evaluation results, including the current study.
13	PG&E	Section 4.2.2 - Permissible Project Types, Page 50	PG&E and DSA, its M&V consultant, recognize the need to have sufficient site screening for the population to be both large enough to get a valid population FSU while also being stringent enough to only admit sites that are good candidates for the program. To state in another way, requirements that are too stringent at the site-level lead to small population NMEC cohorts that may fail FSU requirements. Conversely, allowing too many large and volatile sites runs the risk of having a small number of large, negative savings estimates cause lower-than-forecasted program savings. In fact, in May 2023, PG&E completed an Early Opinion for ED staff entitled, "OBF Prescreening Process/Custom Pathway for Screen Out Projects," which describes the OBF pre-screening process that was developed in 2020 and started in 2021. This document demonstrates PG&E's effort to define eligibility requirements for participation in OBF's Population NMEC cohort and addresses the complexities associated with a diverse population. It's important to also consider OBF's goals as a market support program and balance eligibility requirements with accessibility to an open market of trade pros with varying resources and that serve diverse commercial and public customers. Could the evaluation team offer recommendations as to how to	The Rulebook requirement regarding homogeneity across sites reflects standard practice in consumption data analysis. Lack of sufficient uniformity is why non-residential billing analysis with a comparison group has always been a rare and risky option. There are no recommendations that we are aware of that will change this. The Early Opinion noted PG&E's efforts to address the homogeneity issue based on research by Recurve. Our evaluation shows evidence that the 4% threshold may not be sufficient. It suggests that additional analysis on this is warranted. Presumably, many, if not most, sites could move to site-NMEC-based evaluation, which can address NREs directly. If we have the opportunity to pursue this question, we will certainly start by looking at Recurve's analysis.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
			achieve a level of homogeneity in a non-residential population that balances the need to meet FSU requirements while accurately forecasting savings and maintaining a viable program that supports non-residential customers?	
14	PG&E	Section 4.3.2 - Assessment of PA M&V approaches, Page 58	The Population NMEC Control Group Accuracy Assessment study (CALMAC ID PGE0476.01) conducted several rounds of empirical accuracy assessment testing showing that this method performs as well as other accepted Population NMEC approaches (e.g., CalTRACK + Control), meets FSU criteria, and can perform accurately for a variety of customer segments and industries. The draft report seems to indicate concerns about selection bias, which is always a concern in the absence of randomization, including with methods involving a matched control group. As part of the accuracy assessments performed across residential and commercial populations across California, different matching strategies were tested that attempted to identify the best-performing combination of Granular Profiles (GPs) to include in the final model specification. The final model used reflected that testing (using actual participants in out-of-sample/backcasted tests). Could the evaluation team elaborate on the specific concerns they may have with selection bias with the Synthetic Control plus Time-of-Week Temperature (TOWT) method?	The accuracy assessment conducted by PG&E estimates the accuracy and precision of estimates from various models without program intervention. It indicates that methods based on GPs perform as well as others (including those based on matched comparison groups) in generating accurate estimates that meet FSU precision requirements. However, these tests don't indicate how methods deal with selection bias, which is not observable or measurable. Well-matched groups could result in estimates that reflect self-selection bias to the extent that participants' energy consumption trend differs from the matched groups' trend. We have observed this in practice, where different trends in the component of load not affected by the intervention indicated the presence of selection bias. We are unaware of any readily available approach to test for selection bias. Since it is unclear how well GPs will perform in such circumstances, there is a need for further investigation.
15	PG&E	Section 5 - Conclusions and Recommendations, Page 69	PG&E acknowledges the importance of submitting savings claims in a timely manner (i.e., initial estimates, subsequent true-ups). Under normal conditions, submitting savings claims is straightforward. But in cases of major non-routine events (NREs), as was experienced during the program years 2019 to 2021, submitting savings claims in a timely manner becomes more challenging. To address such challenges, PG&E held discussions with ED staff on how to address them, including ways the process could be accommodated, but to no avail. PG&E would welcome on-going discussions with ED staff and its evaluation consultants on how to improve the claimed savings reporting processes, especially when major NREs occur. Would the evaluators provide a recommended timeline and processes for submitting initial ex-ante savings claims and true-ups to ex-ante savings claims that would serve as a guideline for program administrators to follow moving forward?	Historically, PAs have claimed savings at the end of each quarter following installation. Initial claims for population NMEC programs would continue with this practice. The NMEC Rulebook revision requires that PAs submit quarterly true-ups for initial claims made in CEDARs once 12 months of post-period data become available. Please see the response to comment #3. The evaluation team acknowledges the challenges posed by COVID-19, which coincided with the ramp-up of these programs. The report simply describes what happened. There was no material penalty applied to the PA as a result.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
16	Andres Fergadiotti/SC E-REM	Section 1.1 Study Background, page 1	"*We expect the current NMEC program evaluation process to change in the future, as the NMEC process matures further.*" Question: Based on lessons learned from this study, which specific recommendations on NMEC methods and approach should be considered and/or included in future NMEC rulebook updates?	The evaluation recommended timely program M&V plans and final internal/PA M&V results/reports. These include the timely initial and true-up claims discussed in response to comment #14. The evaluation also recommended a requirement for internal M&V code and data package to be available for evaluation. It also advised the evolution of rules to address challenges related to the precision of program savings, including possible updates to FSU calculations. It also recommended Rulebook updates to address NRE risks, including clarifying program participation eligibility requirements that could cause NREs and for programs to demonstrate that they can sufficiently address these risks. Some of these are already under consideration in the current Rulebook revision. DNV believes these are all changes that will strengthen the ability of population NMEC programs to deliver solid savings, which future revisions should consider.
17	Andres Fergadiotti/SC E-REM	Section 1.3 Study Approach, page 3	**Application of NMEC requirements** – in terms of applicability and with the intent to improved PopNMEC requirements, were NMEC WG's recommendations (dated April 2022) considered and evaluated?	The evaluation considered the April 2022 NMEC WG's recommendations to the extent they were consistent with ongoing CPUC rulebook revisions.
18	Andres Fergadiotti/SC E-REM	Table 1-2. Summary of P4P electric and gas gross savings, PY2019 to PY2021, page 4-5	NMEC Rulebook (C.1 Program Design Criteria) At least 90% confidence / 25% range Fractional Savings Uncertainty (FSU) as calculated using ASHRAE methods at the daily level, or using other methods that achieve at least the same levels of certainty. Question: a. Were the NMEC WG's recommendations (dated April 2022) on FSU evaluated as part of this evaluation? How do WG's recommendation on FSU align with DNV assessment of the methodology? b. Beyond energy savings, how critical was the portfolio size on the FSU estimation for these programs? c. Why did the P4P-Home Energy Rewards program NOT meet the FSU design criteria? Why is the FSU higher (in PY2021)? Were there issues with the savings or size of portfolio? d. In terms of FSU evaluation, how did the portfolio size compare between these programs? Or in different terms, how influential was the portfolio size on the FSU estimates?	a) As indicated above, the evaluation considered the April 2022 NMEC WG's recommendations to the extent that they were consistent with ongoing CPUC Rulebook revisions. b) The level of savings, the size of the program (number of participants), and the homogeneity of participants affect the precision of program savings. In general, size matters for meeting FSU requirements. For P4P programs with a relatively homogenous and high number of participants, even relatively modest savings of about 3% meet the FSU requirement. The OBF program has a relatively low number of participants but fails the FSU requirement even when savings are relatively high (above 6%) because of the heterogeneity of the participating population. c) The poor relative precision results of the P4P-Home Energy Rewards program reflect savings that are small percentages of participant consumption. For a program producing savings this small, meeting 90/25 benchmarks will always be challenging, except with

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
				large populations. The electric savings for this program are small and don't meet the FSU requirement. Only the PY2021 gas savings meet the FS requirement because the size of the program (number of participants) was relatively high. d) As indicated in response to part b of the question, size matters for savings beyond a certain threshold, but the homogeneity of the participating population is also critical.
19	Andres Fergadiotti/SC E-REM	Table 1-5. Population NMEC electric claimed and evaluated savings, PY2019 – PY2021, page 6-7	Gross realization rates for most programs seem to vary significantly from year to year. Assuming methodology accounted for covid effects, variations on bldg. occupancy, solar flags, NREs, etc., why was that the case? Which conditions and/or parameters drove these significant variations in GRRs from year-to-year?	The evaluated gross realization rates (GRRs or our replicated and validated results relative to the most up-to-date PA M&V results) were close to 100% for most programs and years. The GRRs (our replicated and validated results relative to values reported in CEDARS) and their variation are an artifact of the difference between early PA M&V savings results reflected in CEDARS and the final PA M&V savings that DNV evaluated. As the evaluation report indicates, initial claims are reported for each program year, while the trued-up values, negative or positive, are reported in subsequent program years. Final savings are the sum of the initial claimed and trued-up savings reported in CEDARS. However, the revised PA M&V results and the supporting data that DNV received did not match the PA's savings claims in CEDARS. As a result, DNV had to consider which savings numbers to use as the basis for evaluation – the CEDARS claims or the revised PA M&V (we elected to use the PA M&V). As indicated in the report, the evaluated savings for P4P are close to the final PA/internal M&V savings (the evaluation realization rate (in Tables 1-2 and 4-2) and are not variable.
20	Andres Fergadiotti/SC E-REM	Replicability and transparency of PA M&V results, page 9	The Rulebook also says, "All analytical methods, including tools, algorithms, and software used in savings and incentive or **compensation payment** calculations, must be made available to Commission staff and its consultants upon request." (NMEC Rulebook 2.0, III.3.A.2). Question: Why did the impact evaluation excluded analysis and evaluation on methods and approaches leveraged by program implementer to estimate customer's compensation (incentives) payments?	We evaluated the methods used to estimate incentive payments paid to implementers. One P4P program paid customer incentives. As the Rulebook requires, the PA paid 50% of the implementer's incentives based on performance for this program. Analyzing the approach used to estimate customer (end-user) compensation was beyond the scope of the evaluation.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
21	Andres Fergadiotti/SCE-REM	3 METHODOLOGY, page 21-22	**Granular load profile data** – how was the appropriateness by programs on the use of granular load profiles within the NMEC methodology as granular comparison group profiles explanatory variables? How critical were granular load profiles for enhancing adequacy/accuracy of methods? How was this evaluated by DNV? In general, how was the approach for the selection and creation of granular load profiles supported by implementers? Were these processes adequate? Are there specific procedures that can be provided by CPUC/DNV to help the market enhancing NMEC models with the use of granular load profiles?	The first part of the question asks about the appropriateness of granular profiles (GPs) as explanatory variables in the analysis regression model. The PA motivated GPs as explanatory variables in EE regression models (the synthetic control (SC) approach) to address concerns regarding implementer access to individual non-participant hourly consumption data for population NMEC program evaluation. While PA accuracy assessments that investigated the use of SCs to estimate program savings indicate that the approach does as well as other methods, including matched comparison groups, additional studies that analyze its appropriateness are required. One such investigation, for example, could ascertain how well this approach addresses problems that arise due to selection bias (see response to comment #14). As indicated in the report, the evaluation was the tail-end of a long process, which did not leave adequate time for such investigations. We hope to take up such a task in a future evaluation. The second part of the question asks about the process used to create granular profiles and if these are adequate. The implementer constructed GPs as average hourly consumption values of segments (defined by climate zone groups, solar status, load, and premise characteristics) for randomly selected non-participant customers. The construction of granular profiles requires an ongoing technical process to pull random non-participant subsamples, process them correctly, and keep them appropriately updated as the population changes over time. This process and its associated complexity is an additional risk in an already data-intensive process. As a result, DNV has identified further analysis of the granular profiles used for the SC method as an area for additional study. The evaluator hopes to take up this analysis in a future evaluation of population NMEC programs.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
22	Andres Fergadiotti/SCE-REM	3 Methodology, page 21	**Weather data** – how was the appropriateness of the weather data used by NMEC programs methodology evaluated by DNV? Were there any findings and recommendation on this process?	DNV reviewed and ran the PA weather data preparation code, including the weather station mapping used in the PA analysis, and found it appropriate. Both the weather data preparation and mapping follow the methods provided in CalTRACK. The PA M&V process used NOAA weather data and CZ2022 reference temperature files, which provide typical meteorological year (TMY) weather data, for weather normalization. These are standard sources of weather data and are appropriate for the analysis. The process also mapped each participant to the nearest weather station, which is common practice for energy consumption data analysis.
23	Andres Fergadiotti/SCE-REM	3.2.1 Gross savings approach, page 23	**CVRMSE** – "For PY2020, OBF exclusions included those without the required 90% of pre-installation data or for whom baseline models could not be estimated, had poor baseline models (CVRMSE > 1)…" Question: specifically, what was the actual CVRMSE estimated by DNV for each of the evaluated programs' baseline models? Can this be documented in final report? >1? This is too general. ASHRAE recommends for the model not to exceed CVRMSE >0.25. What was the targeted CVRMSE used by DNV on this evaluated to determine adequacy for the baseline models?	The NMEC Rulebook points to a proposed threshold in the LBNL's Site-Level NMEC Technical Guidance of CVRMSE < 0.25 for site-level models. Otherwise, the Rulebook has no other specific CVRMSE requirement, especially for population NMEC programs. CalTRACK recommends 1 (or 100%) as the threshold for building-level models to minimize participant attrition for applications, such as population NMEC, where portfolio-level performance assessment is needed. The PA M&V process used this more permissible CVRMSE threshold, which is reasonable since the focus is on meeting confidence and precision levels at the portfolio level. We now provide the distribution of CVRMSE values for PY2020 and PY2021 P4P and OBF programs in Appendix 6.7 of the report.
24	Andres Fergadiotti/SCE-REM	4.2.5.1 Establishing robust M&V plans, page 49	"… The PA M&V process also excluded P4P sites if they had poor baseline models whose CVRMSE was above a pre-determined threshold (above 1 for the programs under evaluation) …" Question: What is the correct threshold for a model to be considered inadequate? This report references CVRMSE > 1; however, the NMEC rulebook and the latest NMEC's WG guidance references for an acceptable model to have a Coefficient of Variation (Root Mean Squared Error (CV(RMSE)) < 25%. Please clarify. Is there a specific recommendation on what level of coefficient of variation should be used for pre-screening the model? Should projects with a coefficient	Please see the response to comment #23 above.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
			of variation >0.25 or >0.50 be rejected to comply with NMEC requirements?	
25	Andres Fergadiotti/SCE-REM	4.2.5.2 Barriers to applying population NMEC Requirements, page 50	"The PA reported that calculating and meeting FSU requirements has been one of the biggest challenges. FSU calculations for gas and peak savings have been particularly problematic." Per NMEC Rulebook – "At least 90% confidence / 25% range Fractional Savings Uncertainty (FSU) as calculated using ASHRAE methods at the daily level or using other methods that achieve at least the same levels of certainty." $\frac{\Delta E_{save,\,m}}{E_{save,\,m}} = t \cdot \frac{1.26 \cdot RMSE\left(m + \frac{m}{n} + \frac{m}{n}\right)^{1/2}}{m\bar{E}_{pre}F}$ Question: Are models with FSU >> 25%-50% uncertainty still acceptable based on project magnitude, program design, etc.? Are there other validations that needs to be additionally supported e.g., CVRMSE, etc. Are there any specific technical recommendations and/or consideration and/or specific clarifications with methods and approach when estimating the "aggregated" FSU (relative precision) on P4P PopNMEC programs beyond that documented in rulebook?	The Rulebook requirements for FSU in the population NMEC context were unclear. Despite the reference to ASHRAE methods, which provide approaches applicable to site-level models, the clear intent has always been to estimate population NMEC FSU based on aggregate population results (or projections). The Rulebook revisions attempt to clarify this. Aggregate precision estimates are generated based on the formula provided in comment #7. Preliminary FSU calculations (for example, based on engineering-based OBF) will not include the natural variability of consumption data-based savings estimates that the 90/25 FSU threshold is designed to address.
26	Jesse Smith / DSA	1.4.1.1, page 4-5	Did any of the PA claimed savings or PA M&V results have a stipulated NTG ratio embedded? Section 4.2.5.3 indicates that OBF did. Did DNV have to "back out" embedded NTG ratios to create the values in Table 1-2? I'm not suggesting the net results have NTG applied twice, I just think a procedural sentence or two would help. As noted in Section 4.2, some pop-NMEC programs rely on the values stipulated in Section 5.5 of https://docs.cpuc.ca.gov/publisheddocs/published/g000/m232/k459/232459122.pdf. It would be interesting to hear the study team's position on inclusion of stipulated NTGR in PA claims versus withholding analysis of, and adjustment for, attribution for evaluation.	The PA M&V results applied NTG values only to PY2020 OBF estimated savings. DNV did not need to back out NTG ratios to create the values in Table 1-3 (which provide OBF results) since the PA provided M&V results with and without applying the default NTG value for the OBF savings in question. We have included a footnote in Section 4.2.5.3 to clarify this. Based on our understanding, stipulated NTGRs are standard in reporting and do not remove the necessity of ex-post calculation of new NTGRs. Currently, there is no process by which updated, evaluation-based NTGRs replace stipulated NTGRs on a rolling basis, but that would be the appropriate way for the two approaches to be reconciled.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
27	Jesse Smith / DSA	1.4.2, page 8	Regarding “sites removed without explanation”, has DNV received and reviewed the explanations at this point? Upon review were the removals warranted and reasonable? ANC response: They have not yet, but DNV will be getting some additional documentation regarding these removals and why the PA believes they were warranted. In both cases, the reason for exclusion was a large NRE.	DNV has received an explanation from the PA via a comment to the report (see comment # 8). As the response to the question (ANC response) indicates, DNV has not received additional documentation regarding this but expects future PA population NMEC M&V processes to account for the treatment (including removal) of all sites.
28	Jesse Smith / DSA	3.1.1, page 16	Table 3-1 mentions TMY weather data and weather normalization of energy consumption. Are all the results in this report weather-normalized? Or are the results a mixture of avoided energy use and normalized savings? Page 55 seems to have a breakdown. Were incentive payments made based on avoided energy or normalized savings? Where results are weather normalized, a comparison of the avoided energy (not weather normalized) and normalized savings might be informative. I think in the interest of prompt settlement, PA’s might prefer to pay out based on avoided energy. Ultimately, weather normalized is what we want to claim and report. My guess is that there isn’t a lot of exposure for the PA in the weather-normalized savings coming in wildly different from the avoided energy use. The study team’s thoughts on this issue would be a helpful reference point for future implementation and M&V plans. ANC response: Results provided were weather-normalized to TMY data/CZ2020 or CZ2022 normals. This is a continued source of misunderstanding in the popNMEC M&V. Per PG&E’s updated M&V plans, payable savings to implementers are based on ex post or historic weather, while the claimed savings are weather normalized to CZ2022 or the most recent weather normals. As we’ve discussed internally at DSA, this approach allows for implementers to be paid out on a regular basis while ensuring that the claimed savings are not penalized (or over-compensated) by a particularly hot or cold year. This is also the natural input into any program benefit-cost analysis that forecasts EE savings out over the measure lifetime. In most cases, when Recurve or others talk about weather-normalization, they do not mean that they are re-casting savings to a normal weather year. Instead they simply mean they are ‘controlling’ for weather by including weather variables in their M&V regression model.	As the ANC response indicates, payable savings reflect avoided energy use (AEU) savings estimates, and claimed savings are weather-normalized using TMY data. Weather normalizing claimed savings using TMY weather puts results on a long-term weather basis and mitigates ratepayer risk. While AEU (unnormalized) savings estimates allow timely incentive payments, they expose either ratepayers or implementers to risk depending on the extent and direction of baseline and performance period weather differences. Generally, weather-normalization only changes estimates by a few percent. DSA has both the AEU and weather-normalized savings and can assess this. A recognized concern, reflecting recent upward trends in average temperatures, is that the conversion to TMY underestimates cooling and overestimates heating savings, which is a consideration beyond the scope of this evaluation.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
29	Adriana Ciccone / DSA	3.2.1, page 20	Procedurally, it is not very clear that there is a defined process/reporting cadence for program claimed savings, internal M&V, program true-ups and evaluation. I think it would benefit the CA popNMEC stakeholders to produce some additional documentation for how an ideal program evaluation cycle might work. This might include templates for savings reports, recommendations for documenting how FSU or RP was calculated, and the relevant M&V plan components referenced. Something similar to the CA demand response load impact protocols may be a useful template.	DNV agrees that defining initial, internal/PA M&V results, true-up reporting, and evaluation schedules is necessary for CA population NMEC programs. The process could also benefit from reporting templates to facilitate PA reporting and evaluation. For additional information on this topic, please refer to response to comment #3.
30	Jesse Smith / DSA	4.1.1.2, page 34	Could the study team provide a bit more technical detail around the sentence "we calculate precision for population NMEC using the variation of the mean site-level savings across participants"? Perhaps an equation or a recommendation that the next iteration of the NMEC Rulebook provides equations. The report includes a lot of discussion of precision so it's clearly a policy priority. It feels important that everyone is on the same page with respect to how RP and FSU are computed and reported. We've had some internal discussions around the aggregation of errors for peak and net peak kW and how to think about NMEC Rulebook precision requirements for peak and net peak.	Please see response to comment #7.
31	David Jump/kW Engineering	1.4.1.2, page 6	What NTG battery of questions were used to determine the NTG Ratios? Were they adapted for each particular program? Our understanding is that they are set up to inquire about specific measures, not necessarily about a number of measures in the same project. The NTG ratio on the non-residential program is significantly lower than the default value in the NMEC Rulebook. ANC Response: the survey instrument is in the appendix.	DNV has provided the survey instruments in the appendix of the evaluation report. The survey instruments include questions for each program used to estimate NTG specific to each program.
32	Adriana Ciccone / DSA	4.4.2, page 45	I'd like to press on this a bit. Certainly, the OBF sites that were evaluated here did not do well in the popNMEC methods. That, to me, reflects inadequate pre-intervention CVRMSE screening and poor NRE flagging, not necessarily that there was a mix of building types and measures. As I think I mentioned above, DSA and PG&E is looking into this more to understand the best way to screen and group sites to meet FSU and other popNMEC targets. My perspective is that measure types and building types do not necessarily need to be consistent to be part of a popNMEC cohort, assuming that the M&V strategy can demonstrate that it can account for meaningful differences in participants. For example, using industry and climate specific GPs for each OBF participant allow for the same methodology to be used across the population while accepting different participants into the program. As long as the PA can demonstrate that the M&V method meets acceptable targets for both the population and specific subgroups, I don't think that these	The requirement for homogeneity across sites comes from section II.2.B of the Rulebook (provided below). This underlying requirement reflects standard best practice in billing analysis and explains why non-residential billing is relatively rare. The theoretical understanding of how aggregate precision values in a population NMEC context address NREs and provide confidence that they address NREs appropriately needs clarification. It should also be updated to encompass the use of granular profiles in the synthetic control context. Such an undertaking was beyond the feasible scope of the impact evaluation but is a priority for future work. B. Permissible Project Types 1) Site-level rules in this rulebook regarding Permissible

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
			requirements adds much value. This requirement can also limit the population sizes of cohorts such that there will not be a meaningful nonresidential popNMEC program.	Project Types are also applicable to Population-level NMEC Programs. 2) Population-level NMEC program sites must have building-type similarity such that: a. The sites can reasonably be expected to have similar types of equipment holdings, as well as drivers and levels of energy consumption. b. There should be a reasonable expectation that the factors that impact both 1) consumption over a 12-month period, as well as 2) energy savings from program interventions, will be similar across all sites in the population.
33	Adriana Ciccone / DSA	NA	I'd recommend including a set of program definitions for what a program year involves, what first year claimed savings are, and how they should be constructed. We are clear on what the specific sites are that make up the measured savings, but generally confusion in this area should be avoided. Especially as we look at forthcoming evaluations for the MAP programs where installations are spanning multiple years, we should be careful about what 'installation cohorts' represent, what first year savings represent, and how they are different from calendar year savings.	Agreed. Please see the response to comment #2.
34	David Jump/kW Engineering	NA	Will DNV review the comparison group methods used and make recommendations on the advantages and disadvantages of each? Execution of popNMEC M&V requires skilled data analysts on one hand, while more implementers can run SLNMEC M&V analysis. Granular profile methods enable more implementers to capture exogenous effects in their SLNMEC projects. More implementers should equal more projects and savings.	The PA M&V process used granular profiles as synthetic controls to estimate population NMEC program savings. DNV has identified the analysis of this approach to estimate population NMEC program savings as an area requiring further investigation. Please see the response to comment #14 on this.
35	Recurve	General	The draft CPUC's population NMEC evaluation report [IMPACT EVALUATION DRAFT REPORT Population-Based NMEC Programs Program Years 2019 - 2021] by DNV contains a good balance of in-depth computational review (primarily using an audit approach) and complements the impact analysis review with additional metrics and insights that will support program operations and effectiveness in the future. The findings of the report are encouraging. Specifically, the evaluation shows higher savings for the residential population NMEC programs compared to average residential programs, and the high realization rates of the CHR program were highlighted as strong evidence of the effectiveness of targeting.	We agree. The evaluation points to the promise of population NMEC programs. In particular, two of the residential P4P programs (including CHR) motivated more savings compared to similar past programs.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
36	Recurve	Section 1.5	We agree with **requiring up-to-date program implementation plans and program M&V plans prior to evaluation**. We have attempted to assist program administrators maintain up-do-date documentation throughout the course of program implementation and recognize that it is crucial to ensure the transparency, consistency, and accountability of results.	Noted. Thank you.
37	Recurve	Section 1.5	We agree that **requiring a package of internal M&V code and data documented for evaluator replication would be helpful**. It is important to have a well-documented internal system for M&V to facilitate easy auditing by evaluators or other stakeholders.	Noted. Thank you.
38	Recurve	Sections 1.4.1.2 and 3.2.2	**Address concerns regarding free ridership assessment**: The report should critically evaluate the methods used for free ridership assessment, particularly in meter-based programs. Self-report surveys may not be the most reliable method, and alternative approaches should be considered especially when a comparison group is part of the analysis accounts for the majority of the attribution of the program effects. A net-to-gross adjustment is not always necessary when using a comparison group method.	DNV uses well-vetted NTG methods that reflect CA protocols for estimating program attribution. Please see the response to comment #6 on how comparison groups do not result in net savings estimates.
39	Recurve	Section 3.2.1	**Provide more detail on comparison groups and methodologies**: The report should include a more detailed discussion on comparison groups and the methodologies used, as this is a critical aspect of program evaluation.	The implementer constructed granular profiles as average 8760 hourly consumption values of segments for randomly selected non-participant customers. The segments reflected climate zone groups, solar status, premise, and load shape characteristics. Climate zone groups included coastal, inland, North Central Valley, and South Central Valley climate zones. Premise characteristic segments comprised customers with and without electric heat, based on electric rate codes, and consumption quartiles within each climate zone. Load shape characteristics divided customers based on the percent of daily consumption between 1 pm and 4 pm to identify four occupancy patterns within each climate zone, consumption bin, and solar status. The report now includes the above explanation as a footnote in the 5th paragraph of section 3.2.1. Please also note that our report states that this methodology requires further assessment, which did not happen in the scope of the current evaluation.

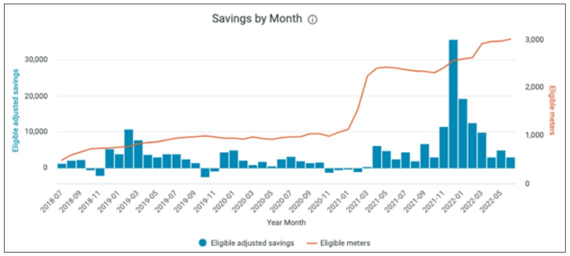
Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
40	Recurve	Section 1.5	**We recommend that the CPUC reconsider FSU requirements for peak electric savings in the NMEC rulebook**. In assessing peak savings, NMEC measurements must balance the need to isolate the small number of true peak hours with the statistical degradation that occurs when isolating a smaller number of hours. If peak kW savings are to be an important metric for the California Energy Efficiency portfolio, we recommend the CPUC collect input and proposals for how to best handle these measurements, including what appropriate FSU bands may be. We'd also suggest that the report acknowledge the benefits of larger portfolios and a strong focus on peak savings as tools for achieving low peak FSU.	DNV's report acknowledges the need to vet peak demand savings estimates for population NMEC programs. Such vetting will require a focus on appropriate FSU or precision for these estimates. We support a process to study and determine a suitable precision level that accounts for the limited number of hours over which peak savings estimates can be made and the portfolio size required to achieve these. We have made edits in the Executive Summary and Conclusion and Recommendations section that acknowledge the need to examine the appropriate FSU or precision for hourly, particularly peak, saving estimates.
41	Recurve	Overall	**Ensure transparency in the removal of outliers from analysis**: Any removal of data points or meters from analysis should be well-documented and justified to maintain transparency and integrity in the evaluation process.	We agree. DNV's report highlights the problem of removing two meters without justification from the PY2021 OBF analysis, which had a material impact on results. The PA has acknowledged this in its reply to this finding. Please see comments #8 and #27 and DNV's response. In general, transparency in how the internal M&V process removes outlier data points or any meters should be part of the analysis method provided in M&V plans and the code used to generate savings estimates.
42	HEA	General	HEA appreciates the opportunity to provide comments on this important document. DNV has accomplished the difficult and important task of providing a comprehensive and independent analysis of California's first P4P residential energy efficiency programs. HEA and PG&E have broken new ground using the P4P model and made many painful discoveries while implementing the innovative and impactful program, HomeIntel.	Noted. Thank you.
43	HEA	Section 4.4.1	**Customer targeting**: From the launch of HomeIntel through 2020 HEA targeted high energy homes with our outreach methods. Any PG&E customer was allowed to participate but outreach material was directed to wealthier communities with larger homes and therefore greater potential savings. This targeted outreach was much more expensive than anticipated, even for a free service. In 2021 PG&E began providing much more cost-effective outreach via email campaigns. Affluent communities were no longer targeted, and the email outreach proved much more successful in increasing participation. As indicated in Table 1.2 of the report, HomeIntel participation increased from 818 in 2020 to 2,648 in 2021. Measured electric savings dropped slightly from 7.7% to 6.7%, but the reduction	DNV's report acknowledges the broadening of participation in the HomeIntel program. The report indicates that the average electric and gas consumption per participant dropped by almost 50% relative to PY2019 levels for the program. It also states that such a drop is consistent with the discussion in section 4.2.1, which indicates that HomeIntel widened its target households after the first year to include participants with more modest consumption levels. Survey responses among participants in the program years under study indicate that, on average, HomeIntel participants are more affluent/educated than those who

<table>
<tr><th>Comment #</th><th>Commenter</th><th>Section</th><th>Comment/feedback/change requested</th><th>Evaluator's Response</th></tr>
<tr><td></td><td></td><td></td><td>in savings was more than offset by the increase in the number participants. With continued PG&E email outreach HomeIntel now has over 15,000 participants and has maintained very high electric savings. In summary, the focus on high energy homes was short-lived, and did not apply to the bulk of participants included in this analysis.</td><td>participated in the other programs, likely and partly a reflection of the earlier customer targeting efforts.</td></tr>
<tr><td>44</td><td>HEA</td><td>Sections 1.4.1.2 and 6.4.3</td><td>How NTGR was determined: DNV defines NTG as “capture the degree to which customer would have installed program equipment or taken program-recommended actions without the program benefits” from page 3 of the report. The definition implies that free or reduced-price equipment will be part of the program. HomeIntel does not provide any equipment or monetary incentives, instead relying on education, targeted recommendations, and ongoing reporting to achieve savings. DNV recognized this and stated on page 6 “HomeIntel is a purely behavioral program that does not incentivize any EE technology. Because of the lack of measure-based savings, we based the program’s NTGR solely on the respondents’ self-reported likelihood of taking the program actions(s)…” However, the question participants responded to was extremely confusing. It assumed the participant had received a “kit”, which would be extremely confusing to them. Below is the survey question from Table 6-3:

Table 6-3. HomeIntel program NTG scoring rubric
<table>
<tr><th>Free-ridership dimension</th><th>Measures applicable</th><th>Question wording</th><th>Answer</th><th>Free-ridership score</th></tr>
<tr><td rowspan="6">Likelihood</td><td rowspan="6">N/A</td><td rowspan="6">Q2. The package of products you received as part of a kit/through the online marketplace the program cost approximately [LOW END PACKAGE]-[HIGH END PACKAGE]. Without the program, how likely would you have been to purchase this performed at your own expense? Would you say…?</td><td>Very Likely</td><td>FR_Liklihood = 1</td></tr>
<tr><td>Somewhat likely</td><td>FR_Liklihood = 0.75</td></tr>
<tr><td>A 50/50 chance</td><td>FR_Liklihood = 0.5</td></tr>
<tr><td>Somewhat unlikely</td><td>FR_Liklihood = 0.25</td></tr>
<tr><td>Very unlikely</td><td>FR_Liklihood = 0</td></tr>
<tr><td>Don't know</td><td>FR_Liklihood = .</td></tr>
</table>
The question starts with “The package of products you received of a kit/through the online marketplace…”. How does a HomeIntel participant answer this question? They did not receive any products and may not have purchased any products on their own. It would be reasonable to answer “Unlikely” to this question since they would probably not buy a kit they never received. The question might have been more appropriately phrased along the lines “Would you have taken energy saving actions without first learning about your energy use and specific suggestions for actions?”. Such a framing would more accurately reflect the program.

But even beyond the poorly phrased survey question is the bigger question of whether NTGR makes sense for HomeIntel. No products or monetary incentives are provided to participants. They take actions based on customized energy education and follow-up. How do you</td><td>The table provides the correct scoring rubric but the wrong question in Question wording column. The correct question asked of HomeIntel participants is "Without the program, how likely is it that you would have taken the action(s)??" The updated table in the revised final report provides the correct question, which was the basis for the program's NTGR calculation.</td></tr>
</table>

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
			quantify whether they would have performed the actions without the education? How would the participant even know? And more importantly, why is this measurement – originally developed for widget-based programs -- an important measurement of the success of a behavioral program? In many ways, HomeIntel is more like the UAT or HERs, other than it provides far superior savings. NTGR of 1.0 is used for these programs (source: CEDARS), and should be used for HomeIntel as well.	
45	HEA	Section 6.4.3	**Concern about customer satisfaction:** HomeIntel received a lower customer satisfaction rating, 68% compared to 79% and 80% for the other 2 programs. CHR has an admirably high satisfaction rating of 80% and it is a high touch program. It is likely that all CHR customers interacted with a contractor, or other program personnel. Such high touch programs should have high ratings but the tradeoff is expense, and getting to scale. Deployment is limited by the number of qualified contractors to install the equipment. HER also received an admirable rating of 79%... but most people will be happy to receive a kit of devices and discounts on purchases. Unfortunately, these incentives didn't lead to significant savings. As indicated in Table 4-22 only 16% of HomeIntel participants interacted with an energy coach. All other customers undertook energy saving measures based on the customized online education and recommendations they received. A 68% overall satisfaction is overwhelmingly achieved through the interactive service. Again, a more interesting comparison would be to compare the customer satisfaction for HomeIntel with UAT or HER, since like these programs, like HomeIntel, are primarily self-service programs.	The Homeintel satisfaction levels were in the middle of the other behavioral programs we have studied or reviewed. DNV's surveys among PY2015 and PY2018 California HER program participants indicated that approximately 45% to 60% reported being satisfied (either very satisfied or satisfied) with the reports in these program years. About 30% to 40% reported being neither satisfied nor dissatisfied. DNV conducted the surveys with PG&E, SCE, SDG&E, and MCE HER recipients. DNV also studied the PY2013-PY2015 impact of the UAT program and conducted surveys among PG&E, SCE, and SDG&E UAT participants. Approximately 45% reported being satisfied, and 45% were neutral. Other HER evaluations have found higher participant satisfaction levels, including DNV's PY2020 evaluation of Puget Sound Energy's HER program and Cadmus' study of Pacific Power Washington's PY2018-PY2019 HER program. These studies found overall customer satisfaction with the HER program of 80% and 85%, respectively. Thus, behavioral programs can also garner relatively high satisfaction.
46	HEA	Section 4.1	**Concern about savings accuracy**: Coby Randolph asked at the end of the 10/18/22 presentation if the report reproduced or validated savings. DNV reproduced the savings methodology used for these 3 program years. This is a significant achievement given the complexity of the problem. Data issues had to be resolved, savings calculations needed to be recreated along with unexpected and extreme complications created by COVID. But reproduction isn't validation. Validation would measure the actual savings seen by participants. Validation would establish the accuracy of the reported savings.	NMEC programs have an internal/PA or embedded M&V process required to pay implementers for energy savings measured at the meter. NMEC programs can choose any M&V methods to estimate program savings that meet the NMEC Rulebook's requirements. In the current study, DNV took steps to determine what an evaluation of programs with an embedded M&V needs to involve. While this needs further clarification, it is unlikely to entail estimating program savings outside the internal/PA M&V process. Even though it was not feasible to examine some critical aspects of the PA M&V methods (such as using

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
				granular profiles as synthetic controls), DNV's effort went beyond simple replication. It involved reviewing the decisions made to get the M&V results, particularly by examining how the PA M&V process defined and handled exclusions, as these can affect estimated savings materially. DNV's evaluation also included an assessment of the M&V approach, including the data preparation and regression methods used to generate the results. Additionally, DNV also calculated relative precisions at the program level to establish the accuracy of estimates. As a first of its kind, DNV's evaluation has helped shed light on how the early population NMEC programs have performed and helped pave the way for future evaluation undertakings.
47	HEA	Sections 4.1 and 4.2	**Concern about savings accuracy**: DNV touches on the issue at the beginning of 4.2.5.2 "…meeting FSU requirements has been one of the biggest challenges. FSU calculation for gas and peak savings have been particularly problematic." In other words, there is low confidence in the reported savings. This aligns with the challenges we've seen for reporting gas savings. The following graph shows the variation in monthly savings reported for HomeIntel's entire portfolio in 2019, 20 and 21:  As shown, gas savings vary tremendously from month to month, even shifting from positive to negative savings and back over a few consecutive months. HEA posits this does not make sense in the real world. As DNV reported, HomeIntel had 2,516 gas sites in 2021. Gas savings across such a large portfolio of homes should not move from positive to negative in a single month. In fact, we would expect smaller and smaller variations month to month as the portfolio increases. This should be irrespective of changes in weather since the CalTRACK methods should normalize such changes. HEA	DNV did not evaluate peak demand savings but believes that methods that address such savings need full vetting. Gas savings estimates were particularly problematic for the OBF program since there were very few participants with gas savings claims, some of whom did not have adequate data. Future evaluations or additional studies should address both of these topics, including the suitability of the estimation methods and reasonable levels of variability for such estimates. As DNV indicated in the report, billing analysis estimates typically reflect data from 50% to 70% of participants due to customer attrition. Average energy savings estimates from customer data included are applied to all program participants since data attrition (primarily due to customers moving) is unrelated to being in the program, and it is reasonable to assume the average savings apply to movers. Because NMEC requires adequate pre-installation data for participation, the proportion of movers was much lower than is typical in billing analysis. DNV assessed that the PA M&V process was rightly conservative in not assigning savings for such movers, given the nature of the program.

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
			believes weather affects are not being effectively normalized by CalTRACK (see explanatory block appended below). HEA concerns regarding CalTRACK's Weather Normalization for Homes We believe month-to-month changes in portfolio-wide savings shown by CalTRACK are inaccurate and are caused by limitations in the models as applied to homes. Justifications: 1. Internal HEA savings calculations show no such widespread variation, neither on a per-customer nor portfolio-wide basis. Weather normalized electric and gas savings are very consistent in our portfolio. • Note: Every month we send each HomeIntel customer a summary of their savings. These figures do not jump up and down. If they did, we would quickly lose credibility with customers who knew they had not made any new changes in each month. 2. A significant portion of our HomeIntel interventions are unrelated to HVAC (and thus unrelated to weather). Most are of the "set and forget" type, which result in consistent savings from month to month. 3. In 2022 our portfolio covered over 6,000 homes. Even if some individual homes showed month-to-month variations there should be minimal variation from month to month across the entire portfolio, due to the law of large numbers. 4. Historically, building analysis best practices have rejected models performing worse than CVRMSE of 0.25 ("poorly behaved buildings"), yet a majority of homes analyzed using CalTRACK are well above this CVRMSE "good buildings" limit. This indicates our modeling needs to be improved. Poor confidence in both gas and peak electric savings should be of great concern. The goal of NMEC programs is to reward savings accurately and at the appropriate times. If gas savings are not being accurately measured, we are not sending the right market signals. HEA argues that gas savings are being undercounted and not correctly rewarded. The analysis performed by DNV does not address this issue. DNV only concludes that gas savings are suspect. Accurate measurement of peak savings may be even more critical than accurate gas savings, since shaving peak energy use is a critical strategy for getting to net zero. Reported electric savings have followed a more consistent course than gas savings, but even there, DNV notes that 10 to 16% of electric sites savings were estimated as opposed to calculated because of very poor CVRMSE values (greater than 1.0; see section 4.1.1.1 page 33).	

Comment #	Commenter	Section	Comment/feedback/change requested	Evaluator's Response
48	HEA	Sections 3.2, 4.3.2, and General	**Concern about savings accuracy**: All these observations point to the shortcomings of the current NMEC methods: savings are modeled using techniques initially developed in the 1990s (PRISM). The landscape has changed significantly since then. Energy use in homes has grown increasingly complex because of the large increases in new electronics and appliances. As we have documented with the NRDC in a 2015 issue paper Home Idle Load: Devices Wasting Huge Amount of Electricity When Not in Active Use, an increasingly large portion of home energy use is now comprised of plug loads. This and other home energy changes prevent homes from being modeled well with current NMEC methods. Our internal analysis over the past decade has shown that high-energy homes are much more difficult to model accurately. This would imply (though we haven't yet verified) that CalTRACK savings calculations for higher energy homes are more frequently inaccurate, due to high plug loads, solar PV, EV charging, and other complex energy profiles. Such homes may increasingly dominate the customer pool. Home energy use has become more complex while data about that energy use has exploded, with the access to granular AMI data. Today this data is underutilized, and NMEC methods need to be improved and validated. PG&E recognized the need to increase the confidence in savings. DNV describes the complexity (section 4.3.2) of introducing the synthetic controls approach to increase the confidence in savings estimates. This approach required significant work and it's unclear if it noticeably increased the accuracy of the results. It does point to the need to find a way to increase accuracy in measuring energy changes. As an industry, we need to be more creative using AMI data to verify energy savings. As far as we know there are no current research projects on how to use AMI data more effectively to analyze home energy. We hope that out of this comprehensive study the CPUC recognizes the need to explore other options for energy savings verification using AMI data. An independent, highly respected organization needs to take up this challenge. Perhaps LBNL, NIST, DOE or other similar organization. We would hope a new method could be developed that is based on the actual energy use in homes and the methodology to perform savings verifications would, like CalTRACK, be an open standard available to everyone in the industry. It would be a standard with national benefits.	We share your desire to have the best methods applied to estimate savings in the population NMEC context and consumption data analysis in general. We are aware of consumption data modeling best practices and the shortcomings inherent in existing methods. A variant of the Time of Week and Temperature (TOWT) model is the basis of the population NMEC program estimates. The specification used can model residential energy consumption and its variability well. However, the synthetic control approach used by the PA here is relatively novel. We have applied some vetting, but at this point, the full range of its strengths and weaknesses is still not fully established. We would be interested in understanding methods HEA employs that improve on other existing methods. How do the savings and precision estimates from such methods differ?