Evaluation Framework for Pennsylvania Act 129 Phase V Energy Efficiency and Conservation Programs

Prepared for the PA PUC

Under Contract 2025-2

June 2026

Demand Side Analytics DATA DRIVEN RESEARCH AND INSIGHTS · NMR Group, Inc. · brightline GROUP · FULCRUM energy solutions

Table of Contents

Figures

Tables

List of Acronyms

ACC: Avoided Costs Calculator	**AEPS**: Alternative Energy Portfolio Standards
AMI: Advanced Metering Infrastructure	**AMR**: Automatic Meter Reading
ARP: Appliance Retirement Program	**ASHRAE**: American Society of Heating, Refrigerating and Air-Conditioning Engineers
ATE: Average Treatment Effect	**B/C** Ratio: Benefit/Cost Ratio
BDR: Behavioral Demand Response	**BER**: Business Energy Report
BRA: Base Residual Auction	**BTUh**: BTU per hour
C&I: Commercial and Industrial	**CaSPM**: California Standard Practice Manual
CBL: Customer Baseline	**CDO**: Commercial Date of Operation
CGEV: Control Group as Explanatory Variable	**CHP**: Combined Heat and Power
CSP: Conservation Service Provider	**Cv**: Coefficient of Variation
CVR: Conservation Voltage Reduction	**DEP**: Department of Environmental Protection
DiD: Difference-in-Differences	**DLC**: Direct Load Control
DLS: Daily Load Shifting	**DR**: Demand Response
DRIPE: Demand Reduction Induced Price Effects	**DDR**: Dispatchable Demand Response
DSM: Demand Side Management	**EAP:** Energy Association of Pennsylvania
EC: Evaluation Contractor	**ECM**: Energy Conservation Measure
EDC: Electric Distribution Company	**EE**: Energy Efficiency
EE&C: Energy Efficiency and Conservation	**EIA**: Energy Information Administration
EM&V: Evaluation, Measurement, and Verification	**EMS**: Energy Management System
EPC: Energy Performance Contract	**ESCO**: Energy Services Company
EUL: Expected Useful Life	**FPC**: Finite Population Correction Factor
FTM: Front-of-the-Meter	**FYSATE**: First-Year Saving Average Treatment Effect
HER: Home Energy Report	**HIM**: High-Impact Measure
HOU: Hours of Use	**HVAC**: Heating, Ventilation, and Air Conditioning
ICSP: Implementation Conservation Service Provider	**IMC**: Incremental Measure Cost
IMP: Interim Measure Protocol	**IPMVP**: International Performance Measurement and Verification Protocol
ISD: In-Service Date	**ISR**: In-Service Rate
kW: Kilowatt	**kWh**: Kilowatt-Hour
LDV: Lagged Dependent Variable	**LED**: Light-Emitting Diode
LFER: Linear Fixed Effects Regression	**LMP**: Locational Marginal Prices
LS: Lagged Seasonal	**M&V**: Measurement and Verification
MEP: Measure-Specific Evaluation Protocol	**MPI**: Market Progress Indicator
MW: Megawatt	**NPV**: Net Present Value
NSRDB: National Solar Radiation Database	**NTG**: Net-to-Gross Savings

NTGR: Net-to-Gross Ratio	**O&M**: Operations and Maintenance
OLS: Ordinary Least Squares	**PDR**: Peak Demand Reduction
PJM: PJM Interconnection, LLC	**PTR**: Peak Time Rebate
PUC: Pennsylvania Public Utility Commission	**PV**: Photovoltaic
PY: Program Year	**RCT**: Randomized Control Trial
RED: Randomized Encouragement Design	**RFP**: Request for Proposal
ROB: Replace on Burnout	**SCADA**: Supervisory Control and Data Acquisition
SO: Spillover	**SSMVP**: Site-Specific M&V Plan
SWE: Statewide Evaluator	**SWE Team**: Statewide Evaluation Team
T&D: Transmission and Distribution	**TOU**: Time-of-Use
TRC: Total Resource Cost Test	**TRM**: Technical Reference Manual
TUS: Bureau of Technical Utility Services	**UEC**: Unit Energy Consumption
UMP: Uniform Methods Project	**VFD**: Variable Frequency Drive
VOI: Value of Information	

Please see Appendix A for a glossary of terms with definitions for many of the items in the table above.

1 Introduction and Purpose of the Evaluation Framework

This Evaluation Framework provides guidelines and expectations for the four Pennsylvania electric distribution companies (EDCs) whose energy efficiency and conservation (EE&C) program plans were approved by the Pennsylvania Public Utility Commission (PUC) to promote the goals of Act 129. The EDCs are Duquesne Light Company, FirstEnergy Pennsylvania, PECO Energy Company, and PPL Electric Utilities Corporation. Through a Request for Proposal (RFP) process, the PUC contracted with a Statewide Evaluation (SWE) Team to complete a comprehensive evaluation of the Act 129 EE&C programs implemented by the four EDCs across Pennsylvania.

To conduct these activities, the SWE Team will collaborate with the four EDCs, their evaluation teams, and PUC staff to develop appropriate, effective, and uniform procedures to ensure that the performance of each EDC's EE&C program is verifiable, reliable, and meets the objectives of Act 129. The SWE Team's tasks include the following:

- Develop the Evaluation Framework, specifying the following:
  - Expectations and technical guidance for evaluation activities
  - Standard data to be collected by implementation conservation service providers (ICSPs) and verified by evaluation contractors (ECs) under contract to the EDCs
  - Audit activities to be conducted by the SWE Team to confirm the accuracy of EDC-reported and EC-verified savings estimates
- Perform ongoing impact and cost-effectiveness audits of each EDC's EE&C plan
- Complete statewide studies and documents, including the following:
  - Periodic updates to the Technical Reference Manual (TRM)
  - Periodic updates to the Total Resource Cost (TRC) Test Order and associated exhibits
  - Statewide Baseline Studies to characterize the market and assess equipment saturation and energy efficiency levels
  - Statewide Market Potential Studies to inform PUC decisions regarding additional cost-effective electric energy efficiency and peak demand savings for a potential Phase VI of the Act 129 programs

The Evaluation Framework is a rulebook that establishes the Act 129 program evaluation process and communicates the expectations of the SWE Team to the EDCs and their ECs. While the document is not a Commission Order and therefore not mandatory, EDCs that align their Evaluation, Measurement, and Verification (EM&V) processes with the Evaluation Framework should expect less scrutiny from the SWE Team as part of the SWE audit activities. The Evaluation Framework outlines the metrics, methodologies, and guidelines for measuring performance by detailing the processes that should be used to evaluate the Act 129 programs sponsored by the EDCs throughout the Commonwealth of Pennsylvania. It also sets the stage for discussions to clarify and interpret the TRM, submit recommendations for additional measures to be included in the TRM, and define guidelines for acceptable measurement protocols for custom measures to mitigate evaluation risks to the EDCs. This requires clear, auditable definitions of kWh/year and kW savings, as well as sound engineering bases for estimating verified gross energy savings.

Specifically, the Evaluation Framework addresses the following:

- Savings protocols
- Metrics and data formats
- Guidance and requirements for claiming reported savings
- Guidance and requirements for gross impact evaluation procedures
- Guidance and requirements for process evaluation procedures
- Guidance and requirements for net-to-gross (NTG) analysis
- Guidance and requirements for cost-effectiveness analysis
- Guidance and requirements for statistics and confidence/precision
- Required reporting formats
- Data management and quality control guidelines and requirements
- Guidance and requirements for data tracking and reporting systems
- SWE Team SharePoint site
- Statewide studies
- Description and schedule of activities the SWE Team will conduct to audit evaluations performed by each EDC's EC and assess individual and collective EDC progress toward the attainment of Act 129 energy and peak demand savings targets
- Criteria the SWE Team will use to review and assess EDC evaluations

Any updates to the Evaluation Framework will clarify and memorialize decisions made through other means, such as Orders, Secretarial Letters, and Guidance Memos. The SWE Team may provide PUC-approved updates as addenda to the Evaluation Framework or by reissuing an updated version of the Evaluation Framework.

1.1 Act 129 Requirements for the Statewide Evaluation

As noted in the introduction, the SWE Team's services include, but are not limited to, the following:

1. Developing an Evaluation Framework
2. Monitoring and verifying EDC data collection
3. Monitoring and assessing the rigor and independence of EDC EC impact and process evaluations
4. Developing and implementing quality-assurance processes
5. Defining performance measures by customer rate class (e.g., sector)

The SWE Team is responsible for auditing the results of each EDC's EE&C plan annually and performing analyses to inform the PUC's updates of overall EE&C program goals for Phase V of Act 129. The audits will include an analysis of each EDC's plan from process, impact, and cost-effectiveness perspectives. The annual audits will include an analysis of plan and program impacts (energy and peak demand savings) and cost-effectiveness. The SWE Team is to report results and provide recommendations for plan and program improvements. The RFP states that the SWE Team will produce an accurate assessment of the potential for energy efficiency, peak demand, and demand response (DR) through market potential assessments. The RFP also specifies that these programs must be implemented pursuant to Act 129 of 2008 and that the evaluations must be conducted within the context of the Phase V Final Implementation Order and Act 129.[1]

As needed, the SWE Team will conduct working groups with the EDCs to encourage improvements to impact and process evaluation techniques. The SWE Team will also produce an accurate assessment of the potential for energy and peak demand savings through a market potential study and provide an analysis with proposed savings targets to inform PUC decisions regarding a possible Phase VI of Act 129. While all these tasks are related, each has distinct goals:

- Impact evaluations seek to quantify the energy, peak demand, and possible non-energy impacts that have resulted from demand-side management (DSM) program operations.
- Process evaluations seek to describe how well those programs operate, characterize the programs' efficiency and effectiveness, and identify opportunities for improvements in program design, marketing, and implementation.
- Cost-effectiveness tests seek to assess whether the avoided monetary cost of supplying electricity is greater than the monetary cost of energy efficiency and conservation measures.
- Market characterizations and assessments seek to determine the attitudes and awareness of market actors, measure market indicators, and identify barriers to market penetration. In addition, they identify current building and equipment stock, standard practices for new buildings and equipment, and they estimate future market direction and practices.

1.2 Roles and Responsibilities

The following tables, adapted from the RFP, delineate the roles and responsibilities for the EDCs, the SWE Team, and the PUC, by tasks and deliverables, per the following categories:

- Statewide Studies

[1] The PUC has been charged by the Pennsylvania General Assembly pursuant to Act 129 of 2008 ("Act 129") with establishing an EE&C program. 66 Pa.C.S. §§ 2806.1 and 2806.2. The EE&C program requires each EDC with at least 100,000 customers to adopt a plan to reduce energy demand and consumption within its service territory. 66 Pa.C.S. § 2806.1. To fulfill this obligation, on June 18, 2025, the PUC entered an Implementation Order at Docket No. M-2025-3052826. As part of the Implementation Order and Act 129, the PUC issued an RFP for a Statewide Evaluator (on October 8, 2025) to evaluate the EDCs' Phase V EE&C programs.

- Audit and Assess EDC Phase V Programs and Results
- Databases
- Primary Data Collection and Impact Analyses
- EDC Plan Review
- Reporting (Semi-Annual and Annual)
- Best Practices
- Other

When appropriate, the SWE Team has classified tasks within the EDCs' primary responsibilities as a role of the ICSP(s) or EC.

Table 1: Roles and Responsibilities – Statewide Studies

Task and/or Deliverable	EDC	SWE	PUC
Conduct electric energy efficiency baseline studies to support the EE Market Potential Study and updates to the Pennsylvania TRM		XX	
Conduct an electric EE Market Potential Study for targets to be achieved in a potential Phase VI EE&C Program from 6/1/31 to 5/31/36		XX	
Conduct an electric Peak Demand Reduction Market Potential Study for targets to be achieved in a potential Phase VI DR Program from 6/1/31 to 5/31/36		XX	
Review and obtain approval of Statewide Baseline and Market Potential Studies (the SWE Team would get approval of these studies from the Commission)			XX
Initiate and coordinate updates to TRM and interim updates (new protocols or corrections)		XX	
Approve TRM updates			XX
Initiate, scope, and conduct/coordinate statewide site inspections, statewide site metering studies, and reviews of data/studies from PA and other states to determine if the PA TRM correctly estimates savings and/or to revise PA TRM protocols (all with focused attention on reducing administrative and overhead costs)		XX	

Table 2: Roles and Responsibilities – Audit and Assessment of EDC Programs and Results

Task and/or Deliverable	EDC	SWE	PUC
Prepare EDC impact and process evaluation plans (EM&V plans), including database and reporting protocols, survey templates, and schedules	EC		
Review and approve the EDC evaluation plans submitted by EDC ECs		XX	XX
Review and update the Evaluation Framework		XX	
Approve the statewide Evaluation Framework and revisions			XX
Conduct impact evaluations, process evaluations, NTG analysis, and cost-effectiveness evaluation	EC		
Review/audit all EDC impact evaluations, process evaluations, NTG analysis, and cost-effectiveness evaluation results		XX	

Table 3: Roles and Responsibilities – Databases

Task and/or Deliverable	EDC	SWE	PUC
Design, implement, and maintain EDC primary program tracking database(s) with project and program data[2]	ICSP		
Establish and implement quality control of EDC program tracking database(s)[3]	EC	XX	
Oversee statewide data management and quality control, including design, implementation, and maintenance of statewide database of program, portfolio, EDC, and statewide energy and demand savings and cost-effectiveness reporting		XX	
Develop and maintain a secure SharePoint site for maintenance and exchange of confidential data and information with EDCs		XX	

Table 4: Roles and Responsibilities – Primary Data Collection and Impact Analyses

Task and/or Deliverable	EDC	SWE	PUC
Collect primary data and site baseline and retrofit equipment information	ICSP/EC		
Perform ex post verification of installation, measure operability, and energy and peak demand savings	EC		
Analyze and document gross and net energy and demand savings at the project, program, and portfolio levels	EC		
Oversee quality control and due diligence, including inspections of project sites, reviews of primary data and analyses, and preparation of claimed and verified savings	ICSP/EC		
Audit and assess the performance of EDC evaluator contractors' EM&V Plans		XX	

Table 5: Roles and Responsibilities – EDC Plan Review

Task and/or Deliverable	EDC	SWE	PUC
Review filed EDC EE&C plans and provide advice to PUC staff on the ability of plans to meet targets cost-effectively (includes cost-effectiveness analyses)		XX	
Review EDCs' EM&V plans and provide advice to PUC staff on the ability of plans to adequately measure energy and peak demand savings		XX	

[2] It is likely that EDCs have internal program tracking database(s). The entry for responsible party is not limited to the ICSP.

[3] It is the ICSPs' and EDCs' primary responsibility for establishing and implementing QA/QC of EDC program tracking database(s). Evaluation contractors should perform QA/QC of an EDC program tracking database. The SWE audits/reviews the QA/QC performed by an EDC, ICSP, and an evaluation contractor.

Table 6: Roles and Responsibilities – Reporting (Semi-Annual and Annual)

Task and/or Deliverable	EDC	SWE	PUC
Report EDC semi-annual and annual energy efficiency and peak demand program and portfolio gross and net impacts, as applicable, as well as cost-effectiveness and EDC progress in reaching targets; conduct process evaluation	EC		
Develop the statewide semi-annual and final annual report templates; review the EDC reports and advise the PUC of program and portfolio results, including results for gross and net impacts, cost-effectiveness, and EDC progress in reaching targets (prepare statewide annual and semi-annual reports for the PUC)		XX	
Review and approve SWE semi-annual and final annual reports			XX
Review the EDC semi-annual and final annual reports and the SWE Team's semi-annual and final annual reports on Act 129 programs. This includes reviewing gross and net savings impacts, cost-effectiveness, and EDC progress in reaching targets		XX	XX

Table 7: Roles and Responsibilities – Best Practices

Task and/or Deliverable	EDC	SWE	PUC
Participate in impact evaluation process reviews and improvements, as needed	ICSP/EC		XX
Prepare best practices recommendations for improvements to impact and process evaluation processes		XX	
Prepare best practices recommendations for program modifications and improvements	EC	XX	

Table 8: Roles and Responsibilities – Other

Task and/or Deliverable	EDC	SWE	PUC
Prepare materials and reports in support of PUC analysis of efficiency programs		XX	
Assist in organizing and conducting periodic stakeholder meetings on evaluation results for energy efficiency and associated Peak Demand Reductions (PDRs), proposed changes to the TRM, etc.		XX	

1.3 Research Objectives

Table 9 displays the Evaluation Framework research objectives for three audiences: the Pennsylvania Legislature, the PUC, and the EDCs.

Table 9: Evaluation Framework Research Objectives

Target Audience	Impact Questions	Process Questions
Pennsylvania Legislature	▪ Did the EDCs meet statutory targets described in Section 2.1 of this Evaluation Framework? ▪ Were energy and peak demand savings calculated via vetted protocols (PA TRM and Evaluation Framework)? ▪ Were the EDC EE&C plans implemented in a cost-effective manner in accordance with the Total Resource Cost (TRC) Test?	▪ Which programs were most successful and why? ▪ Which programs were most cost-effective and why? ▪ If an EDC is behind schedule and is unlikely to meet the statutory targets, how can the EDC improve programs to meet statutory targets?

Target Audience	Impact Questions	Process Questions
Pennsylvania PUC	▪ What level of program energy and peak demand savings was verified for each EDC, and how does this compare to EE&C plan projections and savings reported in EDC semi-annual and final annual reports? ▪ What assumptions related to energy and demand savings need to be updated in future TRM versions? ▪ What were the largest sources of uncertainty related to energy and demand savings and cost-effectiveness identified by EDC evaluators?	▪ Why did planning projections and reported gross savings differ from verified gross savings? ▪ Considering differences in planning projections, reported gross savings, and verified gross savings, how can program planning and reporting be improved? ▪ What actions have the EDCs taken in response to process evaluation recommendations made by the EDCs' EC? ▪ What were the process-related findings of the site inspections conducted by EDC ECs to verify equipment installation?
Pennsylvania EDCs	▪ What factors contributed to differences between planning projections and reported gross savings at the program and portfolio levels? ▪ What factors contributed to differences between *reported* gross savings and *verified* gross savings? ▪ Are there programs or measures that exhibit high free ridership and may warrant a plan revision? ▪ What factors contributed to differences between planned cost-effectiveness and actual cost-effectiveness at the program and portfolio levels? ▪ Which programs require modification or consideration for elimination based on evaluation results?	▪ What changes can the EDCs adopt to minimize differences between planning projections, reported gross savings, and verified gross savings? ▪ What changes can the EDCs adopt to influence customer awareness, satisfaction, and adoption of EE&C programs? ▪ What changes can the EDCs adopt to improve program designs, marketing strategies, and implementation procedures? ▪ Are program participants taking advantage of incentives offered by other program administrators in the Commonwealth? Are ICSPs facilitating coordination with other conservation programs?

2 Policy Requirements

2.1 Requirements From the Phase V Final Implementation Order

Act 129 required the PUC to establish an EE&C program that includes the following characteristics:

- "Require(s) electric distribution companies[4] to adopt and implement cost-effective energy efficiency and conservation (EE&C) plans to reduce energy demand and consumption within the service territory of each electric distribution company (EDC) in this commonwealth"[5]
- Adopts additional incremental reductions in consumption if the benefits of the EE&C Program exceed its costs
- Evaluates the costs and benefits of the Act 129 EE&C programs in Pennsylvania by November 30, 2013, and every five years thereafter
- Includes "an evaluation process, including a process to monitor and verify data collection, quality assurance, and results of each plan and the program"[6]

Based on findings from the Phase V Market Potential Study dated February 2025, the PUC determined that the benefits of a Phase V Act 129 program would exceed its costs, and therefore adopted additional incremental reductions in consumption and peak demand for another EE&C Program term of June 1, 2026 through May 31, 2031 (program years [PYs] 18, 19, 20, 21, and 22). In its Phase V Final Implementation Order, the PUC established targets for reduction in energy consumption and peak demand for each of the four EDCs in Pennsylvania; established the standards each plan must meet; and provided guidance on the procedures to be followed for submittal, review, and approval of all aspects of the EDC EE&C plans for Phase V.[7]

2.1.1 Phase V Energy Reduction Targets for Each EDC

The PUC's June 2025 Phase V Final Implementation Order explained that it was required to establish electric energy consumption reduction compliance targets for Phase V of Act 129. In Phase V, the Commission once again chose to set only peak demand reduction (PDR) targets, rather than dispatchable demand reduction targets, as it did during Phase III. The Phase V Final Implementation Order stated that PDR goals may be met with a combination of long-lasting reductions from energy efficiency or everyday load shifting from demand response programming. The PDR compliance targets represent a single target based on the composite reductions in both the summer and winter seasons.

---

[4] This Act 129 requirement does not apply to an EDC with fewer than 100,000 customers.
[5] See House Bill No. 2200 of the General Assembly of Pennsylvania, An Act Amending Title 66 (Public Utilities) of the Pennsylvania Consolidated Utilities, October 7, 2008, page 50.
[6] See House Bill No. 2200 of the General Assembly of Pennsylvania, An Act Amending Title 66 (Public Utilities) of the Pennsylvania Consolidated Utilities, October 7, 2008, page 51.
[7] Pennsylvania Public Utility Commission, *Energy Efficiency and Conservation Program* Implementation Order, at Docket No. M-2025-3052826, (*Phase V Final Implementation Order*), entered June 18, 2025.

Table 10 contains portfolio budgets, consumption and peak demand reduction targets for Phase V for each of the four EDCs.

Table 10: Act 129 Phase V Five-Year Compliance Targets

EDC	Portfolio Budget Allocation (Million $)	Phase V Consumption Reduction (MWh)	Phase V PDR (MW)
Duquesne Light	$97.7	261,583	46.5
PECO	$427.4	1,111,685	194.8
PPL	$307.5	828,231	151.0
FirstEnergy	$390.3	1,097,605	191.0
Statewide	$1,222.9	3,299,104	583.3

2.1.2 Standards Each EDC's Phase V EE&C Plan Must Meet

The PUC requires that each EDC's plan for Phase V meet several standards, including the following:

1. Each EDC Phase V EE&C Plan must obtain the targeted amount of consumption reduction as stated in Table 11 from programs solely directed at low-income customers or low-income-verified participants in multifamily housing programs. Savings from non-low-income programs, such as general residential programs, will not be counted for compliance. Section 2.1.7 provides additional details about the low-income targets and requirements. Act 129 also includes legislative requirements to include a minimum number of energy efficiency measures for households at or below 150% of the federal poverty income guidelines proportionate to each EDC's total low-income consumption relative to the total energy usage in the service territory. The SWE Team has advised that EDCs should consider the definition of a low-income measure to include a measure that is targeted at low-income customers and is available at no cost to low-income customers.
2. EDCs will be awarded credit for all new, first-year, incremental savings delivered in each year of Phase V.
3. The EDCs' plans should be designed to achieve the most lifetime energy savings per expenditure.
4. EDCs are to develop EE&C Plans that are designed to achieve at least 15% of the target amount in each program year.
5. EDCs are to include at least one comprehensive program for residential customers and at least one comprehensive program for non-residential customers.
6. EDCs should determine the initial mix and proportion of market rate energy efficiency, low-income energy efficiency solar PV, combined heat and power (CHP), and daily load shifting (DLS) programs, subject to PUC approval. The PUC expects the EDCs to provide a reasonable mix of energy efficiency programs and daily load shifting for all customers. However, each EDC's Phase V EE&C Plan must ensure that the utility offers each customer class at least one energy efficiency program.
7. EDCs should report peak demand reduction impacts for summer and winter peak seasons from energy efficiency and daily load shifting programs consistent with this framework.

8. Front-of-the-meter (FTM) measures, such as conservation voltage reduction (CVR) should not make up more than 10% of targeted savings in each EDC's EE&C plan for Phase V.
9. EDCs are encouraged to consider how funding outside of Act 129 from programs such as LIHEAP, DEP's programming, and others may be used when creating their Phase V EE&C plans. EDCs can claim full savings from any EE&C project they support, even though Act 129 is not the sole funding source.
10. EDCs should propose a process in their EE&C plan to help facilitate Alternative Energy Portfolio Standards Act (AEPS) registration for C&I participants of Act 129 programs for their EE projects and take advantage of the elevated alternative energy credit pricing.
11. Phase V EE&C plans should include funding allocated to co-fund American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Level 2 audits or similar technical scoping studies required for participation in other state or federal programs. EDCs can claim savings from any subsequent program participation resulting from the audit.

2.1.3 Carryover Savings from Phase IV

The PUC's June 2025 Implementation Order specifies that gross verified consumption and peak demand reductions that exceed a given EDC's Phase IV target can be applied as carryover toward that same EDC's Phase V compliance targets. Note that only savings achieved in Phase IV can count toward carryover. Carryover savings from Phase III into Phase IV of Act 129 do not factor into the calculation of carryover from Phase IV to Phase V. In response to stakeholder concerns regarding the magnitude of carryover savings in prior phases, the PUC implemented the following limiting mechanisms on carryover savings from Phase IV:

- Energy savings (MWh) are **capped**, meaning that no more than 20% of an EDC's Phase V consumption reduction targets can be achieved by carryover from Phase IV. This applies to both the portfolio consumption reduction targets and the low-income consumption reduction targets.
- Peak demand savings (MW) are **discounted** by a factor of 50%, meaning that EDCs can only carry over half of the peak demand savings they achieve in Phase IV in excess of their Phase IV targets toward Phase V peak demand reduction targets.

FirstEnergy's carryover from Phase IV to Phase V is equal to the net sum of the carryover from the four legacy EDCs: Met-Ed, Penelec, Penn Power, and West Penn Power. FirstEnergy's carryover is subject to the same limiting mechanisms described above.

Carryover of Phase IV peak demand savings into Phase V of Act 129 is permitted despite the differing peak demand definitions across phases. While Phase IV peak demand reduction targets were based on coincident demand reductions from energy efficiency during the summer, Phase V peak demand reduction targets consider both summer and winter peak demand savings. In addition to coincident demand reductions from permanent energy efficiency installations, EDCs can satisfy their Phase V peak demand reduction targets through load shifting programs.

2.1.4 Incremental Annual Accounting

Consistent with prior phases of Act 129, EDCs will be awarded credit for all new, first-year, incremental savings delivered in each year of the phase by their energy efficiency programs. Each program year, the new first-year savings achieved by an EE&C program are added to an EDC's progress toward compliance. Unlike in Phase I and Phase II of Act 129, whether a measure reaches the end of its expected useful life (EUL) before the end of the phase does not impact compliance savings.

The accounting method for daily load shifting programs in Phase V is different from the incremental annual approach described above for energy efficiency and distributed generation. Peak demand reductions from DLS programs are averaged across the phase, rather than summed. This differential accounting addresses the shorter measure life of DLS measures and allows EDCs to claim the resulting MW savings interchangeably toward Phase V goals. The Phase V Final Implementation Order describes the following steps for measuring and claiming peak demand reductions from Phase V load shifting programs:

1. Measure the average reduction in demand over the season-specific peak demand definition.
2. Average the summer and winter performance for a given program year.
3. Compute the average peak demand reduction across the five years of Phase V as the progress toward the peak demand reduction target. Future program years (PYs) are assigned a zero value in this calculation for PY18–PY21 annual reports.

2.1.5 Net-to-Gross Ratio for Phase V of Act 129

The PUC's Phase V Final Implementation Order specifies that compliance will be based on gross verified savings rather than net savings, and that EDCs will continue to perform NTG research. Results of the NTG evaluations should be used to inform program modifications and program planning (e.g., program design, modifying program incentive levels and eligibility requirements), as well as determinations of program cost-effectiveness. Section 3.4 of this Evaluation Framework contains guidance on how EDC ECs should conduct NTG research in Phase V and how the results of this research can be incorporated into program planning and reporting.

2.1.6 Semi-Annual Reporting for Phase V of Act 129

By January 15 of each year, the EDCs are to submit a semi-annual report regarding the first six months of the program year. The format of the report must conform to the semi-annual report template developed by the SWE. Semi-annual reports will only include reported gross savings claims for energy efficiency and distributed generation programs because the gross and net impact evaluations for the program year are not completed by January. Semi-annual reports will include gross verified savings estimates for the summer daily load shifting season because the summer season spans the first three months of each Act 129 program year. Semi-annual reports will not contain reported gross or verified gross savings for the winter daily load shifting season because the winter season concludes after the filing date of the semi-annual reports.

By September 30, the EDCs are to submit a final annual report with verified gross savings for the program year, a cost-effectiveness evaluation (TRC Test), process evaluations, and items required by Act 129 and Commission orders. Section 4.1 provides more details.

2.1.7 Low-Income Customer Savings

As noted in Section 2.1.2, each EDC Phase V EE&C Plan must obtain consumption reduction requirements from programs solely directed at low-income customers or low-income-verified participants in multifamily housing programs (see Table 11 for a summary of the low-income carve-out information). Savings from non-low-income programs, such as general residential programs, will not be counted for compliance. Low-income customers are defined as households at or below 150% of the federal poverty income guidelines. As noted earlier in Section 2.1.3, low-income carryover for Phase V is capped, meaning that no more than 20% of an EDC's Phase V low-income consumption reduction target can be achieved by carryover from Phase IV.

2.1.7.1 Proportionate Number of Measures and Low-Income Savings Targets

Act 129 also includes legislation to ensure that there are specific measures available for and provided to low-income customers. The compliance criteria for this metric are to include a number of energy efficiency measures for households at or below 150% of the federal poverty income guidelines that are proportionate to each EDC's total low-income consumption relative to the total energy usage in the service territory. The SWE Team has advised that EDCs should consider the definition of a low-income measure to include a measure that is targeted to low-income customers and is available at no cost to low-income customers.

Act 129 defines an EE&C measure (in the definitions section; 66 Pa.C.S. 2806.1[m]) as follows:

Energy efficiency and conservation measures.

1. Technologies, management practices, or other measures employed by retail customers that reduce electricity consumption or demand if all of the following apply:
    i. The technology, practice, or other measure is installed on or after the effective date of this section at the location of a retail customer.
    ii. The technology, practice, or other measure reduces consumption of energy or peak load by the retail customer.
    iii. The cost of the acquisition or installation of the measure is directly incurred in whole or in part by the EDC.
2. EE&C measures shall include solar or solar photovoltaic (PV) panels; energy efficient windows and doors; energy efficient lighting, including exit sign retrofit, high bay fluorescent retrofit, and pedestrian and traffic signal conversion; geothermal heating; insulation; air sealing; reflective roof coatings; energy efficient heating and cooling equipment or systems; and energy efficient appliances; and other technologies, practices, or measures approved by the commission.

The SWE Team recommends that EDCs refer to the PA TRM when determining the appropriate level of granularity at which to list measures when calculating the "proportionate number of measures." Technologies that are addressed by a single algorithm section in the TRM should not be further subdivided. Measure divisions should be based on equipment types, not differences in equipment efficiency or sizing of the same type of equipment. For example, EDCs should not separate ductless heat pumps into multiple measures based on capacity. A grouping approach that distinguishes between equipment types but not sizes or efficiency levels should be employed for measures that are not addressed in the PA TRM.

Regarding determining which measures can be classified as specific low-income measures, the legislation states the following:

> *The plan shall include specific energy efficiency measures for households at or below 150% of the federal poverty income guidelines. The number of measures shall be proportionate to those households' share of the total energy usage in the service territory. The electric distribution company shall coordinate measures under this clause with other programs administered by the commission or another federal or state agency. The expenditures of an electric distribution company under this clause shall be in addition to expenditures made under 52 pa. Code ch. 58 (relating to residential low-income usage reduction programs).*

A summary of the low-income carve-out information is provided in Table 11.

Table 11: Act 129 Phase V Low-Income Carve-Out Information

EDC	Proportionate Number of Measures	Phase V Consumption Reduction Target (MWh)	Low-Income Savings Target (MWh)
Duquesne Light	8.40%	261,583	18,933
PECO	8.80%	1,111,685	74,456
PPL	9.95%	828,231	65,678
FirstEnergy	9.33%	1,097,605	86,913
Statewide	**-**	**3,299,104**	**245,980**

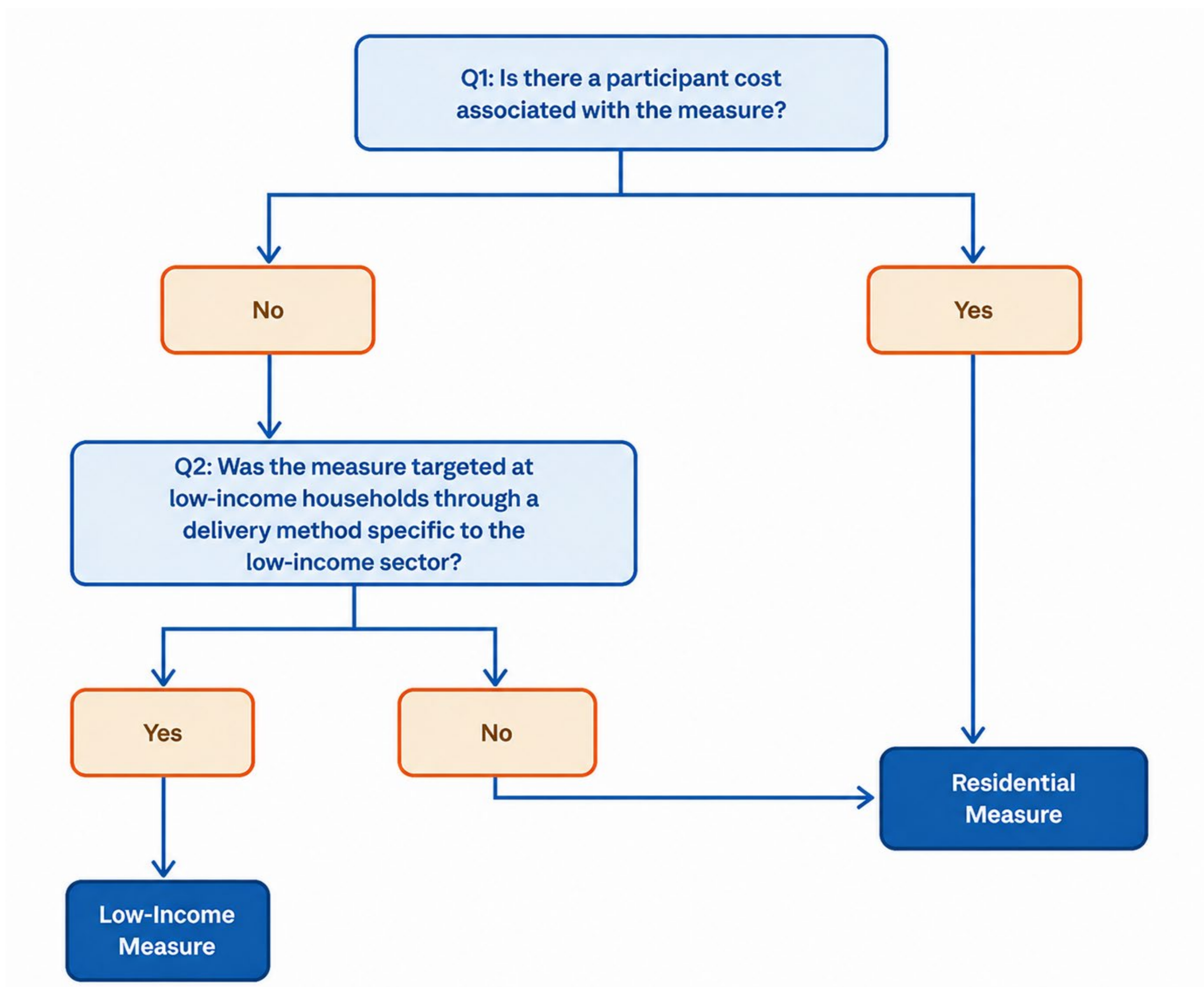
Please note that our recommended definition does not require that the measure/measure type be installed to be counted. Under the definition discussed above, the measure would count if it is targeted to low-income customers and is offered at no cost to low-income customers. If an EDC offers a measure under a specific low-income program (for example, windows) but no customers end up having the measure (windows) installed, it will still count toward satisfying the proportionate measures requirement.

The SWE Team recognizes the possibility of a single measure being classified as both a low-income and a non-low-income measure if it is offered in two different programs with different levels of financial responsibility for the participant. For example, an EDC may offer a heating, ventilation, and air conditioning (HVAC) tune-up measure in its standard residential portfolio where it pays homeowners a $50 rebate toward the cost of the service. The balance of the cost of implementing this measure is the responsibility of the homeowner. This same EDC may offer an HVAC tune-up measure in its low-income program, where 100% of the cost of the improvement is paid by the EDC. In this example, HVAC tune-up should be included twice in the EDC's list of measures offered, but only one occurrence is considered a specific low-income measure.

Figure 1 provides a methodology that EDCs can use to determine whether a given measure in its portfolio is any of the following:

1. A low-income measure (no cost to the participant and targeted to the low-income sector)
2. A general residential measure
3. Offered via two different delivery mechanisms or two different levels of participant cost (free/not free). Therefore, the measure counts once in the numerator of the "proportionate number of measures" ratio and twice in the denominator.

Figure 1: Process Map for Determining Low-Income Measures



During previous Phases of Act 129, several EDCs provided kits to customers in their low-income programs. The SWE believes that each distinct equipment type within these kits should be counted as a separate measure. If an EDC provides low-income program participants with a kit that includes a faucet aerator, a furnace whistle, and a Light-Emitting Diode (LED) nightlight, this should be counted as three measures (faucet aerator, furnace whistle, and LED nightlight) when calculating the proportion of measures offered to the low-income sector.

During Phase III, the SWE identified common practices for differentiating measures for low-income proportionality compliance. The SWE stated these interpretations in Section A.2.3 of the SWE Final Annual Report Act 129 Program Year 11.[8] For Phase V, the SWE recommends EDCs follow similar guidance detailed below:

- **LED Lighting is four measures for residential and seven measures for commercial regardless of bulb or fixture type.** The TRM includes four measure characterizations for residential (*2.1.1 through 2.1.4*) and seven measure characterizations for commercial (*3.1.1 through 3.1.7*). EDCs should not create additional measure variations based on fixture type, location, or lumen output. The analysis used in the SWE's audit combines any more granular measures into one measure per TRM section to be consistent with the SWE recommendation.
- **ENERGY STAR Most efficient refrigerators and ENERGY STAR refrigerators are one measure.** TRM sections, such as *2.4.1 ENERGY STAR* Refrigerators, include two different algorithms that are functionally the same. Both algorithms calculate the difference in efficiency between the old unit and the new unit. The SWE recommends EDCs group these as one measure.
- **Air sealing solutions under TRM section 2.6.1 are one measure.** The TRM has one algorithm section, 2.6.1, that addresses air sealing measures. The main inputs to the algorithm are overall air leakage measurements. The difference in the air leakage measurements is the combined effect of many different air leakage methods (e.g., weather stripping, caulking). In the SWE's analysis, these measures are deemed as part of the Section 2.6.1 algorithm.
- **Advanced power strips count as two measures if EDCs differentiate between Tier 1 and Tier 2 power strips in their reporting.** The TRM has two algorithms for measure 2.5.1 Advanced Power Strips to accommodate two different tiers of smart strip technology. A few EDCs in Phase III only included a single measure for smart strips. If the EDCs provide both Tier 1 and Tier 2 smart strips, then two measures should be counted. If EDCs specify the Tier 1 and Tier 2 measures separately, the analysis conducted by the SWE will count them separately. When EDCs do not specify, the SWE will only count a single Advanced Power Strip measure.
- **Refrigerator and freezer early replacement and recycling should be counted as four separate measures.** The TRM has one section, 2.4.3, that encapsulates all refrigerator and freezer early replacement (replacing an inefficient appliance that has remaining working life with a more efficient model) and recycling (removing an inefficient appliance and preventing it from being used again with or without replacing it). While the TRM does not have different algorithm sections for freezers and refrigerators, the inputs for each are substantially different. Given these differences, the SWE analysis treats them as four

[8] https://www.puc.pa.gov/media/1536/act129-swe_ar_y11_052521.pdf

separate measures. This reflects the difference in benefits generated from replacing an inefficient refrigerator (early replacement) and safely decommissioning an inefficient refrigerator (recycling). In Phase V, the SWE will count 2.4.3 as four separate measures.

EDCs should use the foregoing information as guidance for examining compliance regarding the low-income programs included in their EE&C plans. It is important to note that the proportionate number of measures will be examined when compliance is assessed for Phase V. If an EDC's final annual report shows that there are not enough measures available specifically to the low-income sector, then EDCs will likely be directed to expand their offerings.

2.2 2026 TRC Test Final Order

2.2.1 Intent of the TRC Order

Act 129 of 2008, 66 Pa. C.S. § 2806.1, directs the PUC to use a TRC Test to analyze the benefits and costs of the EE&C plans that certain EDCs must file.[9] The PUC established the TRC Order to provide guidance, methodology, and formulas for evaluating the benefits and costs of the proposed EE&C plans. All cost-effectiveness evaluations and assessments must be conducted in accordance with the TRC Order. The 2026 TRC Test Final Order will be applicable throughout Phase V unless the PUC determines a need to modify the TRC during Phase V.

2.2.2 2026 TRC Test Final Order

The 2026 TRC Test Final Order seeks to provide all instructions for the Phase V benefit-cost analysis of EE&C plans in a single, comprehensive document. The TRC Test Order builds on all previous TRC Test Orders and industry documents, such as the *California Standard Practice Manual – Economic Analysis of Demand-Side Programs and Project*[10] (*CaSPM)*, for the benefit-cost analysis of EE&C plans for Phase V. Updates and refinements to the 2026 TRC Test Final Order include the following:

- Instead of allocating 100% of avoided generation capacity value to summer peak demand reductions, the Commission directed EDCs to allocate avoided generation capacity benefits 50/50 between summer and winter.
- Actual values from Base Residual Auctions (BRAs) for the years that the auctions have been completed should be used for the avoided cost of generation capacity. For all future years, the simple average of the five most recently completed BRAs, escalated using the

[9] The Pennsylvania TRC Test for Phase I was adopted by PUC order at Docket No. M-2009-2108601 on June 23, 2009 (*2009 PA TRC Test Order*). The TRC Test Order for Phase I later was refined in the same docket on August 2, 2011 (2011 PA TRC Test Order). The 2013 TRC Order for Phase II of Act 129 was issued on August 30, 2012. The 2016 TRC Test Order for Phase III of Act 129 was adopted by PUC order at Docket No. M-2015-2468992 on June 11, 2015. The 2021 TRC Test Order for Phase IV of Act 129 was adopted by PUC order at Docket No. M-2019-3006868 on December 19, 2019. The 2026 TRC Test Final Order for Phase V of Act 129 was adopted by PUC order at Docket No. M-2024-3048998 on November 7, 2024.
[10] *The California Standard Practice Manual – Economic Analysis of Demand-Side Programs and Projects*, July 2002, p. 18. Weblink

escalation rate should be used. This is a change from the three-year average used in the prior TRC Test Orders.

- The Commission directed the EDCs to use the EDC- and season-specific avoided Transmission and Distribution (T&D) values calculated by the Phase IV SWE's T&D study.[11]
- The Commission directed the EDCs to include price suppression effects from the Phase IV SWE's Demand Reduction Induced Price Effects study[12] for both energy and capacity in TRC calculations in Phase V.
- The Commission adopted the Phase IV SWE's recommended EDC financial savings benefit from reduced customer arrearages, shutoffs, and collection actions attributable to the EDC's Act 129 low-income programs.[13]

The 2026 TRC Test Final Order specifies that EDCs will continue to use net verified savings in their TRC test for program planning purposes, and cost-effectiveness compliance in Phase V will be determined using gross verified savings.

All EDCs' EE&C plans are required to include both net[14] and gross TRC ratios at the program level separately for EE and DR goals. The 2026 TRC Test Final Order also directed that the Phase V SWE conduct the following analyses for the purpose of reviewing and possibly updating assumptions used for modeling benefits:

- **Vintage of Avoided Cost Forecasts**: compare forecasted avoided costs of electricity to load weighted real-time locational marginal prices (LMPs) and a comparison of forecasted avoided cost generation capacity costs to Base Residual Auction (BRA) clearing prices for each EDC service area.
- **Compliance with Alternative Energy Portfolio Standards Act (AEPS)**: summarize the AEPS costs in the Phase V SWE final annual reports and identify any significant differences between the forecasted and actual AEPS costs.
- **Societal Benefits**: further investigate the validity of the potentially reduced costs of providing rate assistance to Customer Assistance Program participants attributable to EDC low-income programs regarding the appropriateness and magnitude of such a benefit in future TRC Test Orders.

2.2.3 Avoided Cost Calculator

Consistent with Phase IV, the Commission maintained the Act 129 methodology to develop avoided cost forecasts of electricity through use of a MS-Excel spreadsheet calculation model (Avoided Costs Calculator or ACC) developed by the SWE. The objective of an established calculator is to provide more

---

[11] Exhibit 4 to the 2026 TRC Test Final Order: The Avoided Cost of Transmission and Distribution Capacity Study. Weblink

[12] Exhibit 3 to the 2026 TRC Test Final Order. DRIPE Study. Weblink

[13] Exhibit 2 to the 2026 TRC Test Final Order. Arrearages Study. Weblink

[14] The PUC's Phase V Final Implementation Order required the inclusion of net TRC ratios, in addition to gross. See Phase V Final Implementation Order at page 222.

nuanced instructions and guidance, which should improve consistency across EDCs and better align with market conditions. The Commission, in its 2026 TRC Test Final Order, directed EDCs to use this standard tool when developing avoided costs for Phase V.[15]

2.2.4 TRC Order Schedule

The PUC issued a Final Order for the TRC Test for Phase V of the Act 129 EE&C program on November 7, 2024, and determined that the 2026 TRC Test Final Order shall apply for the entirety of Phase V. Reviews will be undertaken when warranted, and changes will be made only when justified during a phase. The PUC determined that it is necessary to keep the TRC parameters constant to compare the actual Phase V benefits and costs to the planned Phase V benefits and costs using a definition of TRC costs and benefits that remain constant over Phase V.

2.3 PA TRM Order and TRM Manual

In implementing the AEPS Act, 73 P.S. §§ 1648.1 – 1648.8, the PUC adopted energy efficiency and DSM Rules for Pennsylvania's AEPS, including a TRM for Pennsylvania on October 3, 2005.[16] The PUC also directed the Bureau of Conservation, Economics, and Energy Planning[17] to oversee the implementation, maintenance, and periodic updating of the TRM.[18]

Like Phase IV of the Act 129 EE&C program, the PUC adopted the 2026 TRM as a component of the EE&C Program evaluation process for Phase V.[19] The TRM Order represents the PUC's continuing efforts to establish a comprehensive and up-to-date TRM with a purpose of supporting the EE&C Program provisions of Act 129. The PUC will continue to use the TRM to help fulfill the evaluation process requirements contained in the Act. By maintaining up-to-date information, the PUC assures that Act 129 monies collected from ratepayers are reflecting reasonably accurate savings estimates.

The 2026 TRM is organized into three volumes. The first volume provides guidance and overarching rules regarding the use of the TRM. The second volume contains TRM protocols, or measure-specific methodologies for estimating energy and demand savings, for residential measures. The third volume contains TRM protocols for commercial and industrial (C&I) and agricultural measures. The TRM also contains appendices to present information that does not easily fit the template of a TRM protocol.

---

[15] Exhibit 1 to the 2026 TRC Test Final Order. 2026 Avoided Cost Calculator. Weblink
[16] Order entered on October 3, 2005, at Docket No. M-00051865 (October 3, 2005 Order).
[17] As of August 11, 2011, the Bureau of Conservation, Economics, and Energy Planning was eliminated and its functions and staff transferred to the newly created Bureau of Technical Utility Services (TUS). See Implementation of Act 129 of 2008; Organization of Bureaus and Offices, Final Procedural Order, entered August 11, 2011, at Docket No. M-2008-2071852, at page 4.
[18] See October 3, 2005 Order at page 13.
[19] Current and prior versions of the TRM are posted on the PA ACT 129 TRM webpage. Weblink

2.3.1 Purposes of the TRM

The TRM serves a variety of purposes for Act 129. In addition to providing measure savings protocols, the TRM ultimately seeks to facilitate the implementation and evaluation of Act 129 programs. The TRM fulfills the following objectives:

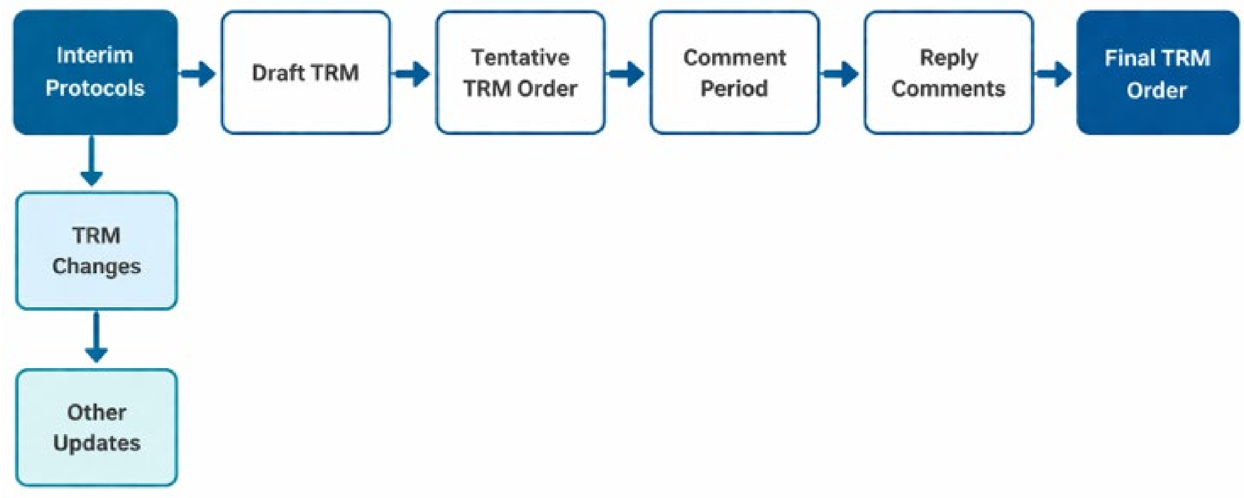
- Serves as a common reference document for energy efficiency measures to be used by EDCs, ICSPs, ECs, the SWE, the PUC, and other stakeholders.
- Establishes standardized, statewide protocols to calculate energy and demand savings for measures. The ICSPs use these protocols to estimate ex ante (reported or claimed) savings achieved for the energy efficiency measures. EDC ECs use these protocols to estimate ex post (verified) savings achieved for energy efficiency measures.
- Increases transparency to all parties by documenting underlying assumptions and tracking references used to develop savings estimates for measures.
- Balances the accuracy of savings estimates with costs incurred to measure and verify the savings estimates.
- Provides reasonable methods for measurement and verification (M&V) of incremental energy savings associated with EE&C measures without unduly burdening EDC EE&C program implementation and evaluation staff.
- Reduces the number of EE&C measures that must be evaluated as custom measures.

2.3.2 TRM Update Process

In Phase III, the PUC made the 2016 TRM effective for the entirety of the Phase but reserved the right to implement a mid-phase TRM update as deemed necessary. For Phase V, like Phase IV the PUC will use the 2026 TRM and reserves the same right to implement a mid-phase change.[20] All changes made during the TRM update process will be prospective and thus will not retrospectively affect savings determinations for the program year already underway, unless otherwise determined by the PUC. Updates to the TRM will occur per the typical stakeholder process, which adheres to the Tentative Order, Comment Period, Reply Comments, and Final Order procedure (see Figure 2).

[20] *Phase V Final Implementation Order*, p. 208.

Figure 2: TRM Update Process



Any party may request the addition of a new savings protocol to be added to the TRM. The Bureau of Technical Utility Services (TUS) staff, along with the SWE, the EDCs, and their evaluators, will help to review, clarify, and/or improve new and existing savings protocols. Protocols for any measures that are not already included in the TRM may be proposed through the Interim Measure Process (Section 2.3.5).

As impact evaluation results become available and changes to federal and state energy codes and standards are implemented, they will serve as indicators to identify measure protocols that may require updates in the TRM. A review process will explore the applicability of these findings to ensure that the TRM presents the best available estimates of energy and demand savings. Measure attributes will be updated through dedicated measure research studies informed by the impact evaluation findings during the review process.

For Phase V, as in Phase IV, the PUC adopted a process for allowing optional updates to keep the TRM aligned with updates to codes and standards that occur during the phase. Each year of the phase, the SWE will track updates to federal standards, ENERGY STAR specifications, and state-adopted building energy codes. Based on the extent of code updates that occur, the SWE will recommend whether to open the TRM for a code refresh for the following program year. Code updates that are not finalized and in effect before July 1 of a program year will not be considered for inclusion in the TRM in that update cycle. Changes to the TRM proposed by the SWE through this process will be limited to updating values directly related to codes, standards, and ENERGY STAR specifications. Any modification to the Phase IV TRM will become effective on June 1 of the calendar year following the comment and review process (Table 12). In place of a formal TRM Order and update, the PUC may also elect to direct the SWE to develop and issue a series of guidance memos summarizing the changes and recommending that the EDCs update their savings calculations and reported savings to reflect the changes. Codes and standards memos will be posted to the SWE SharePoint site rather than the Commission's Act 129 website.

Table 12: Timeline for Process for Code Change Updates

Estimated Date	Action
March 15	SWE memo analyzing impact of code or standards changes will be delivered to TUS.
April 15	TUS will determine if an update is warranted.
July 1	Codes and standards must be in effect by this date.
July	Tentative TRM Order and Manual on Public Meeting Agenda.
August – September	Comment and review process.
November	Final TRM Order and Manual on Public Meeting Agenda.
June 1	Code Change Updates become effective

The SWE encourages EDC ECs to recommend improved coincidence factor values using a load shape from metered or vetted sources to estimate peak demand impacts. The SWE will consider the proposed values for prospective TRM updates. The SWE reserves the right to request additional documentation to investigate the applicability of the load shapes submitted.

2.3.3 TRM Protocols

A TRM protocol, or measure characterization, is a measure-specific methodology for calculating energy and demand savings. The TRM contains protocols that determine savings for standard measures by either deeming savings or providing an algorithm with variables to calculate savings. Protocols to estimate energy and demand savings associated with behavioral modification programs, daily load shifting programs, and conservation voltage reduction are included in Section 6 of this Framework.

The Pennsylvania TRM categorizes all non-custom measures into two categories: deemed measures and partially deemed measures.

- *Deemed measures* are well-defined measures that have specified (fully stipulated) energy and demand savings values; no additional measurement or calculations are required to determine deemed savings. For some deemed measures, deemed savings may be provided for various permutations of the measure in a lookup table. For example, the C&I Refrigeration Auto Closers measure provides separate deemed savings values based on space type (walk-in cooler or walk-in freezer) and climate zone.
- *Partially deemed measures* are determined using an algorithm with stipulated and open variables, thereby requiring data collection of certain parameters to calculate energy and demand savings. Partially deemed measure protocols may include "default savings" tables that list the energy and demand savings values calculated using defaults for all open variables. Unlike deemed savings, the use of default savings values is optional. For example, the ENERGY STAR Ceiling Fan measure provides default savings for various product types and fan diameters. The default savings values rely on a daily hours of use assumption of three hours. EDCs have the option to replace this default value based on primary data collection and then use the algorithms provided to calculate savings.

Custom measures are considered too complex or unique (because there are highly variable or uncertain savings for the same measure) to be included in the list of standard measures provided in the TRM and so are outside the scope of the TRM (Section 2.3.3.3).

2.3.3.1 Deemed Measures

A deemed measure protocol specifies a pre-determined amount of energy and demand savings per unit. For the PA TRM, deemed measure protocols also may contain an algorithm with stipulated variables to provide transparency into deemed savings values and to facilitate the updating of the deemed savings values. Stipulated variables, which are assumptions that must be used and are established through the TRM update process, cannot be changed mid-cycle without approval from the PUC.

This type of protocol typically is used for measures whose parameters are well understood or well documented. It is particularly appropriate for residential measures involving customers with similar electricity usage characteristics, as well as for kit programs.

SWE recommendations to the PUC regarding TRM deemed savings protocols for future years include the following:

- Maintain an active TRM working group, chaired by the SWE, including technical experts from the utilities and other independent experts to provide input on evolving technologies and measure assumptions.
- Identify measure protocols to be reviewed in the phase based on relative savings contributions, evaluation findings, statewide studies, changes to federal and state energy efficiency codes, and recent secondary research.
- Conduct a periodic review of national deemed savings databases to determine how others have used this tool and the assumptions they have utilized.
- During the TRM update process, examine literature referenced in the TRM that supports the deemed savings assumptions; this would include reviewing the population or tests from which the data were derived and recommendations about the population or technologies to which the generalizations should be applied in Pennsylvania.
- Update the TRM measures to reflect changes in federal and state energy efficiency codes and standards.
- Update the TRM to address findings of the program evaluations.

2.3.3.2 Partially Deemed Measures

The Pennsylvania EE&C programs include several measures that utilize savings measurement protocols based on partially deemed savings. Customer and equipment-specific information is used for each open variable, resulting in a variety of savings values for the same measure. This method is commonly used when well-understood variables affect the savings and can be collected from the applicant, distributor, or retail channel partner. TRM Appendix C (Lighting Audit & Design Tool for Commercial and Industrial Projects) and Appendix D (Motor & Variable Frequency Drive [VFD] Audit and Design Tool) are provided to facilitate the calculation of energy savings and demand reductions for partially deemed lighting, motor, and VFD measures.

It is noteworthy that measures proposed within Midstream and Upstream programs are mostly characterized as partially deemed, as savings are likely to vary based on capacity, efficiency, and/or configuration along with customer location weather characteristics. Some open variables may have a default value to use when the open variable cannot be measured. Examples of open variables include the following:

- Capacity of an A/C unit
- Lumen output of an LED fixture
- Square footage of insulation
- Hours of operation of a facility or of a specific electric end-use
- Horsepower of a fan or pump motor

SWE recommendations to the PUC regarding TRM partially deemed savings protocols for future years include the following:

- Identify high-impact measure (HIM) protocols for review and providing necessary clarifications or modifications based on evaluation findings, statewide studies, changes to federal and state energy efficiency codes, or more recent and reliable secondary research available.
- Conduct Pennsylvania-specific research studies to update key assumptions for HIMs and provide hourly load shapes for each measure variant.
- Analyze algorithms and definitions of terms during the TRM update process to verify that the protocols use accepted industry standards and reasonably estimate savings.
- Analyze low-impact measures with unrealistic and inaccurate savings values. Review low-impact measures periodically to adjust the level of EM&V rigor based on market adoption.
- Ensure that the methodologies for implementing protocols are clearly defined and can be implemented practically and effectively.
- Establish energy impact thresholds for non-residential measures in the TRM, above which customer-specific data collection is required for open variables. The intent of this change is to reduce the overall uncertainty of portfolio savings estimates by increasing the accuracy of project-level savings estimates for extremely HIM installations.
- Add new measures and associated algorithms for increased industry prevalence of end-use equipment controls, consumer electronics, and connected equipment (Smart Devices).

2.3.3.3 Custom Measures

The TRM presents some information about custom measures that are too complex or unique to be included on the list of standard measures in the TRM. Accordingly, savings for custom measures are determined through a custom measure-specific process, which is not contained in the TRM (see Section 2.3.6).

2.3.4 Using the TRM

The TRM provides a standardized statewide methodology for calculating energy and demand savings. The TRM also provides a consistent framework for ICSPs to estimate ex ante (claimed) savings and for EDC ECs to estimate ex post (verified) savings.

2.3.4.1 Using the TRM to Determine Ex Ante Savings

This section outlines how ICSPs should calculate ex ante savings.[21]

For replacements and retrofits, ICSPs will use the applicable date to determine which TRM version to select to estimate EDC claimed savings.[22] The installation date or *commercial date of operation* (CDO) should be the date at which the measure is installed and energized.

For projects with commissioning, the CDO is the date commissioning is completed, and the equipment is installed and energized.

For new construction, selection of the appropriate TRM must be based on the date when the building/construction permit was issued (or the date construction starts, if no permit is required) because that aligns with codes and standards that define the baseline. Savings may be claimed toward compliance goals only after the project's installation date. For projects that overlap phases, the TRM in effect on the date the permit was issued should be selected regardless of which phase the project was completed in.

Methods used by the ICSPs to estimate ex ante savings differ for each of the three measure categories (deemed, partially deemed, and custom measures).

For **deemed measures**, ex ante savings are determined by applying the deemed savings values in the TRM. Assumptions, which may be listed in the TRM for transparency, may not be adjusted by ICSPs using customer-specific or program-specific information.

For **partially deemed measures**, ex ante savings are determined by using the algorithms provided in the TRM; these formulas include both stipulated and open variables. Stipulated variables are defined as any variable in the TRM that does not have an EDC Data Gathering option. These values may not be changed or revised by ICSPs. Open variables in the TRM have an *EDC Data Gathering* option. These values can come from either customer-specific information or default values provided in the TRM. ICSPs should attempt to collect customer-specific values for each rebated measure through the application process. Only variables specifically identified as open variables may be adjusted using customer-specific information. If the ICSPs choose to utilize the EDC data gathering option for a particular open variable, the findings of the EDC data gathering should be used for all instances of that variable. ICSPs are not allowed to revert to the default value once the EDC data gathering option is chosen. However, if customers or midstream market actors are unable to provide data for the variable, then ICSPs should use the default value found in the TRM for those customers only. For measures where EDC data gathering is utilized, EDCs should report on findings in their final annual reports.

---

[21] In some cases, an EDC may choose to implement a program *in-house* rather than engaging an implementation CSP. In these cases, EDC staff is acting in the capacity of the implementation CSP.

[22] Pennsylvania Public Utility Commission Act 129 Phase II Order, Docket Nos.: M-2012-2289411 and M-2008-2069887, Adopted August 2, 2012, language in Section K.1.b. Commercially operable is defined as the equipment is installed and energized.

The SWE will collaborate with the EDCs and their evaluators during the TRM update process to identify any stipulated variable that should be changed to an open variable and vice versa. The criteria for making such changes may include the feasibility of attaining such information, the percentage change in savings expected when using open versus stipulated variables, and the uncertainty surrounding default values.

For certain non-residential end-use categories, the TRM defines thresholds where M&V is required if the threshold is exceeded. In other words, if the combined kWh savings for a certain end-use category in a single project are above the corresponding end-use category threshold established in the TRM, the ICSP cannot use default values but is instead required to use customer-specific data collected through M&V activities. If claimed savings for an end-use category (e.g., lighting, motors) within a project fall below the threshold specified in the TRM, the ICSPs may gather customer-specific data or use the default TRM value.

It is helpful for ICSPs to use the same approach as the EC for determining when they must use customer-specific data gathering to estimate ex ante savings. EDCs or ECs should assist the ICSPs in interpreting the requirements of this Evaluation Framework, including determination of ex ante savings methodologies at the project and/or measure level. The use of similar methodologies to estimate savings between the implementers and evaluators will increase the likelihood of a strong correlation between ex ante and ex post savings and improve the precision of savings estimates for a given sample size.

For **custom measures**, ex ante savings are determined using the custom measure process described in Section 2.3.6.

Measures that are not included in the TRM but still require a deemed or partially deemed approach may be claimed using the Interim Measure Protocol (IMP) approach described in Section 2.3.5.

2.3.4.2 Using the TRM to Determine Ex Post Savings

Typically, EDC ECs field surveys, conduct site inspections, and perform documentation reviews on statistically representative samples to determine ex post savings. The appropriate method used to determine verified savings differs for the three measure categories and may further depend on the magnitude of the project's savings. These measure categories, defined below and summarized in Table 13, dictate the methodology to use for estimating ex post savings.

Table 13: Measure Categories

Measure Category	Ex Post Calculation Methodology	Example Measures
Deemed measures	Follow deemed savings per TRM	Air handler filter whistle
Partially deemed measures	Follow TRM savings algorithms, using deemed variables and verified open variables	C&I lighting, residential HVAC, C&I motors
Custom measures	Follow Behavioral protocol (Section 6.1), applicable UMP protocol or other custom measure protocols developed for the project	Behavioral Programs, Non-TRM compressed air equipment, non-TRM chiller, Energy Management System (EMS)

For **deemed measures**, the TRM provides per-unit savings allowances that both the ICSPs and evaluators will use. The energy and demand savings of these measures are deemed with all energy-related variables stipulated. Thus, the evaluation activity for deemed measures will include verification of measure installation, quantity, and compliance with the eligibility guidelines of the TRM measure protocol. The evaluator will estimate ex post savings using deemed savings and/or stipulated assumptions in accordance with the TRM.

For **partially deemed measures**, the EDC EC will estimate ex post savings using the algorithms provided in the TRM. These formulas include both stipulated and open variables. The open variables typically represent or describe straightforward, key measure-specific inputs in the savings algorithms that improve the reliability of savings estimates (e.g., capacity, efficiency ratings). Evaluation activities for partially deemed measures include verification of measure installation, quantity, and the correct use of the TRM protocol; verification of open variables, which may entail confirming nameplate data; facility staff interviews; or measurements of the variable(s). Evaluators should attempt to verify as many open[23] values in the TRM algorithm as possible with customer-specific or program-specific information gathered through evaluation efforts. Open variables in the TRM may have a default stipulated value, which should be used if customer-specific or program-specific information is unreliable or the evaluators cannot obtain the information.

Customer-specific data collection and engineering analysis will depend on the type of measure (uncertainty and complexity) and the expected savings (level of impact). The ICSP is primarily responsible for collecting customer-specific data through supporting documentation, phone or in-person interviews with an appropriate site contact, a site visit, pre- and post-installation metering, analysis of consumption histories, analysis of data from building monitoring equipment, and/or energy modeling simulations. For example, estimating savings for commercial lighting projects requires detailed information about pre- and post-installation conditions for lighting retrofits, such as fixture and ballast type, fixture wattage, building and space type, hours of use (HOU), and lighting controls. When required by the TRM, using more accurate customer-specific values for a partially deemed measure is mandatory for high-value non-residential projects above a threshold kWh/year.[24] ECs should

[23] Open variables are signified by the term "EDC data gathering" in the TRM.
[24] The threshold kWh/year is stipulated in the TRM and will vary depending on the type of measure.

verify the customer-specific data for all measures in sampled projects above the threshold. If the EC determines that the customer-specific data gathered by the ICSP is not reasonably valid, then the evaluator should conduct independent customer-specific data gathering activities for those measures. A Site-Specific Measurement and Verification Plan (SSMVP) is required for all projects with combined measure savings above the TRM thresholds.

Section 3.3.2.3 provides additional information on non-residential savings thresholds for project stratification and determination of measure-level rigor.

For **custom measures**, the savings impacts vary per project. The customer, the customer's representative, or a program administrator typically estimates the project's savings before an EDC pays the incentive. Due to the complexity of custom measures and the information required to reasonably estimate savings for them, EDCs may choose how to estimate reported gross savings. The EDC EC must verify reported gross savings to an acceptable degree and level of rigor. In some cases, evaluation activities may require the measurement of energy and/or demand consumption, both before and after the implementation of the custom measure. In other cases, engineering models and regression analysis may be permitted. Therefore, the audit activities for custom measures typically depend on the evaluation process selected for the category of custom projects.

2.3.4.3 Using Off TRM Protocols to Determine Savings

For both deemed measures and partially deemed measures, if an EDC wishes to report savings using methods other than the applicable TRM, they may use a custom method to calculate and report savings, as long as they (1) alert the SWE to the planned departure in their evaluation plan, (2) calculate the savings using TRM protocols, and (3) include both sets of results in the EDC reports. The EDCs must explain the custom methods in the final annual reports, wherein they report the deviations. If an EDC uses a custom method to calculate savings for a TRM measure, the SWE will only perform a pre-approval review if the PUC requires them to do so.

Custom methods to calculate savings differ from using program-specific or customer-specific information for open variables defined in the TRM protocols (see Section 2.3.4.1).

2.3.5 Interim Measure Protocols

IMPs are used for measures that do not exist in the TRM and for additions that expand the applicability of an existing protocol. IMPs serve as a holding ground before a protocol is fully integrated into the TRM.

The SWE will maintain a catalog of IMPs, showing their effective dates on the SWE Team SharePoint site, to maintain a database for new/revised measure protocols that should be included in subsequent TRM updates, for EDCs to use to claim ex ante savings, and for evaluators to follow when determining ex post savings.

2.3.5.1 Interim Protocol Approval Process

The SWE classifies each IMP as either a *new measure interim protocol*, an *existing measure expansion interim protocol*, or a *TRM modification interim protocol*. IMPs classified as new measure interim protocols or existing measure expansion interim protocols undergo an informal approval process. The approval process is intended to minimize risk for EDCs planning to offer measures that do not have a TRM protocol by developing savings protocols through a collaborative review process with the SWE. IMPs classified as TRM modification interim protocols are not reviewed by the SWE.

For a new measure or existing measure expansion interim protocol, the IMP review and approval process includes the following steps:

1. EDCs submit IMPs to the SWE.
2. The SWE reviews a proposed IMP and returns any suggested revisions to the submitting EDC.
3. After discussion and revision, the SWE sends the IMP to the other EDCs for comment.
4. After an IMP undergoes an iterative review process between the SWE and the EDCs, the SWE gives the protocol interim approval as an "approved interim measure protocol."
5. Interim approval is formalized when the SWE confirms approval via email and posts the final protocol and its effective date on the SWE Team SharePoint site. The approved protocol is available for use by all EDCs.
6. The SWE includes all IMPs in the next TRM update for public comment and review and formal approval by the PUC.

New Measure and Existing Measure Expansion Interim Protocols

This category of interim protocols refers to completely new measures or additions that expand the applicability of an existing protocol, provided that the additions do not change the existing TRM algorithms, assumptions, and deemed savings values. For new measures and expansions of existing measures, an approved IMP will apply for the remainder of the phase in which it was approved. The IMP, whether changed or unchanged, will apply prospectively; an IMP will not apply retrospectively, unless the PUC formally approves a request to do so.

TRM Modification Interim Protocols

This category of interim protocols refers to EDC-proposed modifications to existing TRM protocols. This category includes proposed changes to an existing TRM protocol that modify the existing TRM algorithm, assumptions, and/or deemed savings values. Modifications to existing measures are normally performed during the PUC-approved TRM update process, but EDCs can propose TRM modifications of critical importance between TRM updates. Any EDC-developed TRM modification to interim protocols must be provided to the SWE for informative purposes. However, neither the SWE nor Commission staff will review and approve the protocol. If an EDC uses such a protocol, that EDC will report savings using both the existing TRM protocol and the modification protocol. The TRM modification interim protocol may be used to inform the next TRM update.

2.3.6 Custom Measures

While TRM measures are reviewed and approved by the PUC through the TRM update process, custom measures do not undergo the same approval process. This section describes a process for managing custom measures by establishing a method for documenting energy and demand savings; describing the general requirements for custom measures; and clarifying the roles of the EDC, ICSP, EC, and SWE Team.

EDCs may report ex ante savings for a custom measure according to methodologies used by the customers or contractors and approved by the ICSP. EDCs are not required to submit ex ante savings protocols for custom measures for SWE approval. ICSPs must perform measurements consistent with IPMVP options to collect baseline and/or post-retrofit information for custom measures that have estimated savings above a threshold kWh/year level.[25] ICSPs are encouraged to perform measurements for custom measures with estimated savings below the threshold. To reduce the likelihood of significant differences between ex ante and ex post savings, ECs are encouraged to recommend the IPMVP option and M&V protocols to be used by the ICSP.

The PUC will not determine M&V protocols for custom measures in order to improve the EDCs' ability to support energy services that meet the EDCs' energy savings goals. ECs are permitted to determine the appropriate M&V protocols for each project. ECs must verify impacts for custom measures selected in the impact evaluation sample. They must develop an appropriate SSMVP for each sampled project, according to their professional judgment. SSMVPs should be uploaded to the SWE Team SharePoint site two weeks before the site inspection is scheduled by the EC. ECs must verify the project-specific M&V data (including pre- and post-metering results) obtained by the ICSPs, as practicable, for projects in the evaluation sample.

If the EC determines that data collected by the ICSPs are not reasonably valid or do not cover an adequate range of operating conditions, then the evaluator must perform measurements consistent with IPMVP options to collect post-retrofit information for custom measures that have estimated savings above a threshold kWh/year level. The EC must make baseline assessments in the most efficient and cost-effective manner, without compromising the level of rigor. It is strongly recommended that ICSPs reach out to ECs to ensure that baseline assessments are conducted in an acceptable manner, baseline definitions are representative of what would have happened in the absence of the measure(s), and that all necessary data points are collected for the estimation of savings. SSMVPs should include a detailed discussion of the baseline, including how it was determined and why it is feasible from the home/business owner's technical and economic standpoint. This baseline feasibility discussion should at a minimum address the following elements, as applicable:

1. Capable of meeting the relevant needs of the facility – e.g., production quantities and speeds, cooling capacities, compressed air volume, etc.

[25] TRM savings thresholds should also be used for custom measures.

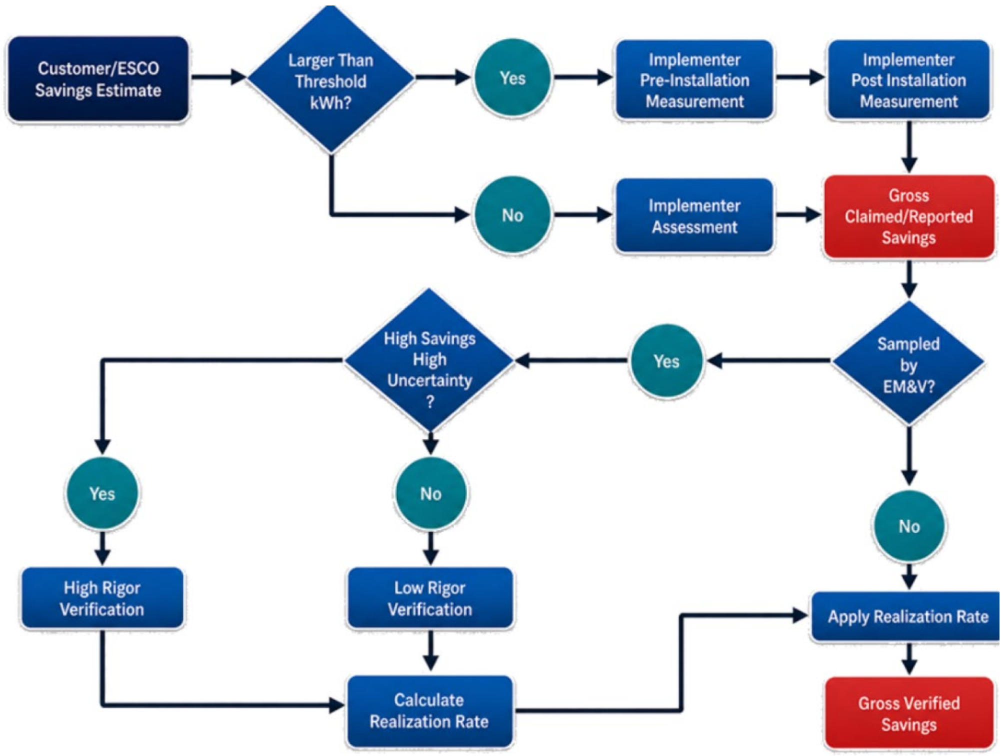
2. Equipment is readily available for purchase in new condition
3. Not considered a temporary solution
4. Offers a level of service comparable to the efficient measure and/or market standards

For early replacement baselines, ECs must work with ICSPs to verify the viability or functionality of the existing equipment. Evidence of equipment viability may include demonstration of existing equipment physical operation. This supporting documentation is expected to be collected as the project is being developed and should not be created or re-created after the fact. If no documentation supporting the baseline condition is provided to the EC, the EC is expected to establish an appropriate baseline based on the best available information, prioritizing hard evidence that is objective, quantifiable, and verifiable, (e.g., photos, contractor removal quotes, performance data, etc.) over soft evidence that may be subjective and could include interpretation (e.g., interviews). Further, in the event custom projects are submitted without proper baseline condition documentation, the EC is expected to discuss updating the program's baseline condition documentation requirements and/or enforcement of those requirements with the ICSP.

The SWE reserves the right to audit, and review claimed and verified impacts of any custom measures or projects. The SWE will randomly choose projects sampled by the ECs and will audit the evaluators' engineering analysis and realization rates. In addition, the SWE also may select a random sample of projects not sampled by the EDC ECs and conduct an independent assessment of the ex post savings. The SWE may use these independent samples to augment the sample selected by the ECs. The results from SWE independent assessments may be included in the program's realization rate calculations at the discretion of the EC. Figure 3 presents a flow chart of the generic process to verify savings for custom measures. Deviations from the process are acceptable.[26]

[26] For example, not all projects above the kWh/year threshold will require baseline measurements. Some may require only post-retrofit measurement.

Figure 3: Custom Measure Process Flow Chart



Only observable savings from M&V may be claimed or verified. For projects where savings are expected to grow after the M&V period and are therefore not reflected in the M&V results, EDCs and their ICSPs are encouraged to claim additional savings as a second phase of the project at the time those additional savings are also observable from M&V. For example, in a manufacturing facility that is only operating the program-supported high-efficiency equipment at 60% capacity at the time of post-installation M&V, only savings observed during the M&V period may be claimed or verified. Additional savings from increased production may be claimed or verified when M&V can observe those savings.

An exception to allowing only observable savings from M&V is when there is extensive documentation and/or evidence supporting future savings growth of a project. The SWE reserves the right to review and/or reject proposed future savings based on inadequate documentation and/or evidence. If claiming higher future savings, ICSPs and ECs should average future expected savings for the remainder of the Act 129 phase. For example, a wastewater treatment plant may claim savings from adding efficient pumps and VFDs beyond the plant's current load if the project can prove that the plant's load will grow over time due to documented housing and business growth in the area it serves. Industrial production or sales growth projections from businesses are generally not considered sufficient evidence for supporting higher future savings estimates, unless the sales or production growth projections are supported by firm documentation such as customer purchase orders or plans to discontinue production at another facility.

For projects that incur a disproportionately higher material and/or installation labor cost initially but are scheduled to produce savings incrementally over time, the cost allocated to the project should be staged according to reasonable expectations of savings achievement over time. This cost staging plan should be completed at project initiation. Later project stages should have separate program enrollments to allow for accurate accounting across Program Years and Act 129 Phases. If later project stages do not materialize, EDCs should still claim the remainder of the project cost already spent but not already claimed in previous project stages.

For example, if a 1,000 kW CHP plant is installed and fully operational but its electrical production is planned to ramp up over two phases at 300 kW and 700 kW (due to interconnection issues, on-site operational coordination, etc.), the proportion of the CHP plant's total cost allocated to the 300 kW phase should be 300 kW / (300 kW + 700 kW) = 30%. The remaining 70% of the CHP's total cost should be allocated to the second 700 kW phase. If the second phase does not materialize, the EDC should still claim the remaining 70% project cost for TRC calculations under a second program enrollment.

2.4 Guidance Memos

The SWE Team developed this Evaluation Framework to provide an overarching framework for Act 129 programs and therefore may not address all nuances discovered through the actual implementation and evaluation process. For such issues, the SWE will develop guidance memos to clarify and memorialize decisions through an iterative review process with input from EDCs, their ECs, and TUS staff. These guidance memos will be the last step in resolving open issues and will formalize high-level decisions that impact all EDCs.

The SWE will post all PUC-approved guidance memos with their effective dates to the Phase V SWE Team SharePoint site. Guidance memos issued by the SWE in prior phases have been incorporated into this Evaluation Framework, as appropriate. Neither guidance memos nor SWE documents or positions necessarily reflect the opinions, regulations, or rulings of the PUC and, therefore, are not binding on the PUC. On an annual basis, the SWE will review and retire any guidance memos that become obsolete.

2.5 Study Memos

It may be necessary to conduct evaluation-related research studies to support the program design or evaluation analysis efforts. Study memos outline a specific research topic for the SWE to investigate. The SWE will work with the EDC teams to identify the need for any near-term and long-term research studies. These collaborative efforts will minimize redundant, independent research and reduce costs. TUS staff will be responsible for approving any SWE-conducted research studies. The SWE will primarily collaborate with EDCs through collection of data from previous implementation and evaluation activities. TUS staff are responsible for approval of study memos. Results from these studies are intended to inform updates of the TRM and TRC Test or support planning for future phases of Act 129. As the research studies are identified and approved for implementation, all activities will be completed under existing budgets, unless otherwise noted. The SWE will distribute study memos to EDCs for information purposes.

3 Technical Guidance on EM&V

This section of the Evaluation Framework is intended to help guide EDC ECs in the development and execution of successful evaluation plans. Section 3.1 contains the SWE's recommendations and requirements for evaluation plan development. Each efficiency measure that is implemented as part of an EDC's EE&C plan is assigned a reported (ex ante) impact estimate for energy and demand savings. These ex ante savings values are usually generated by an ICSP retained by an EDC to administer a specific EE&C program and associated efficiency measures. Determination of the ex ante savings values are based primarily on TRM protocols; this is discussed in Section 3.2.

The sum of the savings reported (through program tracking databases and systems) by the EDC and/or its ICSP is the gross reported savings for the EE&C program. However, compliance with Act 129 savings targets is based on gross verified savings estimates. To develop these estimates for a program, an EDC's EC selects a sample of projects from the program population for verification of the ex ante savings estimate, which may include more rigorous M&V activities than those used to prepare the reported savings estimates. These M&V activities are discussed in Section 3.3.

A sample is typically used because it is not feasible or cost-effective to evaluate each of the hundreds or thousands of efficiency measures implemented. Section 3.6 presents the annual evaluation sampling requirements at the portfolio, sector, and program level, and offers technical guidance on sample design, allocation of resources, and presentation of the uncertainty introduced by sampling on gross verified impacts. Section 3.6.4 describes other sources of uncertainty in an evaluation and how ECs should address these factors.

3.1 EDC Evaluation Plans

Planning is a critical first step in successful program evaluation. The evaluation plan, or EM&V plan, outlines the approaches the evaluator will use and serves as a guiding document for the evaluation. EDCs must complete an initial evaluation plan for each program and submit it to the SWE Team SharePoint site for review within 151 days of the start date of Phase V (by October 30). The evaluation plan should be a single electronic document that includes, at a minimum, sample design, frequency and schedule of evaluations, and the high-level M&V approach. It should contain a chapter for each program in the portfolio, or a separate document for each program. Final evaluation plans are due December 15 of PY18 (December 15, 2026).

Within four weeks of submission of the draft Phase V EM&V plan, the SWE Team will either approve the plan or suggest modifications to it. If the SWE Team suggests modifications, the EDCs will have two weeks to submit revisions based on the SWE comments and submit a revised evaluation plan. Then the SWE Team will have two weeks to provide final comments or approve the revised plan. Either party may request a time extension if unforeseen circumstances arise.

Changes to program delivery and evaluation approaches can occur from one year to the next within a program phase. The SWE Team recommends that EDCs submit a redline version of the evaluation plan for Program Years 19–22, or whenever intra-year changes are required. The SWE will attempt to provide an expedited review of updated evaluation plans and either approve the plan or suggest modifications to the revised plans within two weeks of submission. ECs are encouraged to submit evaluation plan modifications to the SWE as early as possible in the program year.

Each EDC and its EC will choose the optimal structure and design for their evaluation plans. The evaluation plan should at least reflect a shared understanding of the program delivery mechanisms, research objectives and methodology, data collection techniques, site inspection plans, and intended outcomes. Evaluators should discuss the gross impact evaluation, NTG analysis, process evaluation, and cost-effectiveness evaluation activities and outcomes separately. Evaluation plans should also contain a proposed timeline of activities and a table of key program contacts. Evaluation plans should identify who will conduct site inspections (the EDC, the ICSP, the EDC's EC, or some other entity) and the type of site inspections (in-person or virtual). Evaluations plans should also explain how the EDCs would make site inspections results available to the SWE Team. Sections 3.3 through Section 3.9 provide technical guidance to the EDC ECs regarding evaluation plans and activities for Phase V of Act 129.

The PA TRM provides EDCs with open variables for numerous energy conservation measures (ECM savings parameters). Often, a default value is provided as an alternative to customer-specific or program-specific data collection. An EDC evaluation plan should identify open variables for which the ICSP or EC intends to utilize the option of EDC data gathering. The SWE encourages the EDC evaluators to utilize as many open values in the TRM algorithms as possible with customer-specific or program-specific information, particularly if the data are gathered by ICSPs and tracked in EDC data tracking systems. For parameters where the ICSP uses TRM default values for efficiency or simplicity, evaluators are encouraged to use the EDC data gathering option for cases in the evaluation sample. The SWE expects the results of these data collection efforts to be used in the calculation of verified gross savings, even if the resulting savings differ from the impacts calculated from using the default value.

3.2 Reported Savings

3.2.1 Tracking Systems

For the EDC ECs to evaluate programs, it is imperative that EDCs maintain complete and consistent tracking systems for all Act 129 programs. The tracking systems should contain a central repository of transactions recorded by the various implementation ICSPs capable of reporting ex ante savings. The values in the tracking system should be used for reporting ex ante energy and demand savings, customer counts, and rebate amounts in the EDC semi-annual reports. EDC tracking systems must also be capable of fulfilling the SWE's standardized quarterly data request, as described in Section 4.2.1. Home Energy Report (HER) and Daily Load Shifting programs are excluded from the quarterly data request. Records stored in EDC tracking systems also should be the basis of the EC's sample selection processes and contain project parameters relevant to the savings calculation for each installed measure.

The SWE should be able to replicate summations from the tracking systems and match the summed savings value for a program and initiatives within a program, sector, and portfolio to the corresponding values in the EDC semi-annual and final annual reports. EDCs must ensure that the tracking system contains all of the fields that are required to support calculation and reporting of program ex ante savings.[27]

3.2.2 Installed Dates, Recorded Dates, and Rebate Dates

An EDC tracking system must capture several important dates:

- **Installed Date**: The date at which the measure is physically installed and operable. For upstream rebate programs, such as lighting or appliance programs, it is appropriate to use the transaction date as the installed date since the actual installation date is unknown. For new construction projects, the installed date is the date the equipment is energized even if the building is not yet occupied or will not be used until another, unrelated installation/project is completed.
- **Recorded Date**: The date the measure is entered into the program system of record for future reporting to the PUC. This does not refer to the submission date of a semi-annual or final annual report.
- **Rebate Date**: The date the program administrator issues a rebate to the participant for implementing an energy efficiency measure; this may be substituted with an Approval Date, which is the date a rebate is approved for payment within an implementer's system, if there is a time delay between approval of a payment and issuance of the rebate/incentive.
- **Filed Date**: The date an EDC officially submits and files a semi-annual or final annual report to the PUC as part of a compliance requirement.

Reporting lag occurs when the savings for a transaction are reported in a later quarter/year than the quarter/year the measure went in-service. For example, a measure may go in-service in PY18 but not be recorded or reported until PY19. There are two types of reporting lags:

- *Participant lag* describes the time between when a participant buys and installs a measure and submits the associated rebate application to the program administrator; this can be as brief as a few days or as long as six months. This lag largely depends on participant behavior and program policies.[28]
- *Approval lag* describes the time between when a customer submits a rebate application and the program administrator approves the application; this will vary by program and project, and stems from key program processes, such as application review, QA/QC procedures, installation verification, and rebate and invoice processing. Approvals of program transactions are guided by EDC communications related to eligibility and deadlines for

[27] Some worksheets used in the calculation of individual customer impacts will not be embedded in the tracking system but can be provided upon request.

[28] Act 129 and Orders approving programs recognize savings for measures installed after a specified date. Different programs and program managers may have policies and communications that can impact customer lag.

program application submittal. Similar processes exist for upstream buy-down programs that require time for retailers and manufacturers to compile finalized sales documentation.

The SWE has defined a process for dealing with the two types of reporting lag as related to reporting to the PUC. EDCs are directed to file final annual reports by September 30 following the end of the program year[29] (i.e., 120 days after the end of the program year), which works well for projects with installation dates prior to the end of the program year but recorded dates following the end of the program year. In tandem with their final annual reports, EDCs may submit Q5 measure tracking data. Though there is no fifth quarter to the program year, the Q5 tracking data will include measures that were not in prior tracking data submissions due to reporting lag.

In rare cases where the recorded date follows the final annual report deadline, but the installation date is prior to the end of the program year, EDCs must provide a supplemental report with the final verified savings of lagged transactions by the semi-annual reporting deadline (January 15) of the program year following the measure's installation date. For the PY22 final annual report / Phase V report, in addition to the supplemental reporting in the PY23 semi-annual report, EDCs must include an appendix in the PY22 annual report that describes the timing challenges, provides an inventory of the late projects by program and estimates of the associated reported gross energy and peak demand savings. The PY23 semi-annual report addendum should provide the verified gross savings of the end of Phase V projects and include an updated and final estimate of carryover savings from Phase V into Phase VI (assuming there is carryover from Phase V to Phase VI). Verified savings should be estimated by applying the relevant PY22 realization rate to the late projects, unless the late project is above the metering thresholds, as defined in Table 1-2 of Volume 1 of the 2026 TRM, in which case the SWE-approved metered savings should be considered verified. The addendum does not need to include updated estimates of net savings, NTG ratios, or cost-effectiveness results and TRC values.

This assumes the PUC will establish a Phase VI energy efficiency and conservation program of Act 129. If a Phase VI is not established, the SWE will provide an update to the guidance on end of phase reporting for projects not included in the PY22 final annual report due to lags in project recording and documentation.

Situations may arise in which it is unclear what is the appropriate TRM or IMP to use for savings calculations. The SWE and TUS staff agreed that the applicable date for determining which TRM to use (for all measures, excluding new construction) is the installation date. TUS staff and the SWE concluded that the installation date is the correct date to use because it marks the date when the customer starts to realize savings and ensures that savings calculations match the date when they begin to accrue. ICSPs and ECs should use the TRM in effect at the installation date when calculating energy and demand savings for Phase V. For new construction, selection of the appropriate TRM must be based on the date when the building/construction permit was issued (or the date construction starts if no permit is required) because that aligns with codes and standards that define the baseline. Savings may be claimed toward compliance goals only after the project's installation date. This requirement is to

---

[29] *Phase V Final Implementation Order*, pp. 210–211

account for the long lifecycle of new construction projects that are designed to a particular standard prior to construction.

3.2.3 Historic Adjustments

EDCs are required to document any adjustments made to ex ante savings after a semi-annual or final annual report and quarterly data request response has been submitted. Any change to the reported kWh impact, reported kW impact, or rebate amount for a claimed project is considered a historic adjustment. The SWE understands that such adjustments must be made to correct errors, or reflect better information, but requires that the EDC inform the SWE of these historic adjustments prior to the submission of the EDC's final annual report. This process will allow the SWE to update its records and track program progress using the corrected values. Two acceptable methods for submitting these historic adjustments are as follows:

1. **Record replacement** – This technique involves submitting two new records for the measure being revised. The first record will be the inverse of the original tracking record submitted to the SWE (negative kWh, kW, and incentive amounts) and will serve to zero out the original values submitted. The second record should contain the corrected project impacts.
2. **Record revision** – This technique involves submitting a single record containing the adjustments to project parameters. For example, if the original measure record contained an impact of 1,300 kWh and it was later discovered that the correct gross reported savings value for that measure is 1,650 kWh, the new tracking record would contain a reported kWh value of 350 kWh.

With either approach, the EDCs should identify historic adjustments using an indicator variable set equal to 1 for an adjustment record and equal to 0 for a new tracking record. This indicator variable is needed to produce accurate participation counts by quarter or program year because a project receiving historic adjustments should not be included when determining the participation count for the program (because it was counted previously). If an EDC has an alternate methodology for informing the SWE of historic adjustments to ex ante impacts that is not listed in this section, the approach can be submitted to the SWE Team for consideration and approval.

3.2.4 Key Fields for Evaluation

Because the EDC evaluators use equations to independently calculate verified savings for some partially deemed TRM measures, the SWE requires that the EDCs capture and provide key variables used to calculate savings to the EDC evaluator. The EDC's ICSP should collect these variables so the evaluator will not have to retrieve the variables independently for projects outside of the evaluation sample. For projects in the evaluation sample, it is the EC's responsibility to independently verify each parameter in the savings calculation.

3.2.4.1 Key Data Collection Fields for Energy Assessments or Audits

Some program delivery models include an audit or assessment of homes or businesses to identify energy saving opportunities. These audit or assessment reports developed by the ICSP contain essential data for program evaluation and should be collected with care, rigor, and consistency. An audit report shall be completed for each participant/unit on a standard form. At a minimum, the following information should be included for each participant/unit:

- Participant characteristics (name, address, account number, premise number, phone, etc.)
  - If multifamily, ideally provide information on landlord/property manager and on individual tenants in units served
- Vendor providing services
- Existing home characteristics, such as conditioned square footage, space heating fuel, water heating fuel, number of occupants, and premise type
- List of individual measures implemented within the measure group, such as AC replacement, AC maintenance, number of LEDs, refrigerator removal, refrigerator replacement, faucet aerator, showerhead, water heater pipe insulation, water heater tank insulation, water heater replacement, attic insulation, blower door guided air sealing, duct wrap, etc.
- Denotation of whether service is provided at a single- or multifamily residence
  - If multifamily, the number of units served
  - If multifamily, denotation of measure installation by unit
  - If multifamily, denotation of measures installed in common areas
- Details on individual measures, such as the following:
  - Existing lamp and replacement LED wattage, and room where the LED is installed
  - Existing and replacement heat pump capacity, model number, efficiencies, etc.
  - Existing and replacement refrigerator type, model number, configuration, etc.
  - Number of faucet aerators and showerheads and water heating fuel
  - Replacement insulation R-values
  - Estimated deemed or engineering-derived energy savings per unit installed
  - Estimated savings for all measures installed at a particular account

3.3 Gross Impact Evaluation

3.3.1 Overview

This section establishes guidelines for all ECs that conduct gross impact evaluations. Impact evaluations determine program-specific changes, which include reductions in electric energy consumption, peak demand savings, changes in fossil fuel or water consumption, and avoided greenhouse gas emissions.[30]

---

[30] While EDCs are not required to report emissions impacts in EE&C program impact evaluations, estimates of greenhouse gas avoidance can easily be estimated based on verified gross energy savings and emissions factors from sources such as PJM, the Energy Information Administration (EIA), and the Federal Energy Regulatory Commission.

As there are many stages to an impact evaluation, decisions must be made at each stage based on the desired accuracy and certainty of the evaluation results and the funds available. Section 3.3 provides evaluators information to support decision-making throughout the gross impact evaluation process.

Evaluators leverage the data collected during program implementation and conduct independent data-gathering activities. If the data collected by the ICSP are unreliable, if end-use equipment operating conditions have changed post-installation, or if the ICSP did not conduct or complete project-specific data collection activities for a project with high informational value, the EC(s) must collect the appropriate data for sampled projects. A primary responsibility of the EC is to independently validate that the measures are installed as claimed and that the equipment counts, capacities, efficiencies, and other key specifications recorded by the ICSP are accurate for a statistically representative sample of program-supported equipment. Simply recalculating savings based on the values stored in the program tracking data is a useful exercise, but it is not an independent evaluation.

The EM&V activities may include surveys, reviews of project documentation, or direct observation and measurement of equipment performance and operation at a sample of participant sites to verify that the energy savings reported for the projects are correct and that the equipment is installed and operating. Successful impact evaluations assess the costs incurred with the Value of Information (VOI) received and balance the level of evaluation detail (rigor, as defined in Section 3.3.2.2) with the level of effort required (cost). How deeply an evaluator goes into the assessment of key variables at a sampled site or among program participants depends on the value of that information in verifying the reported savings.

Approved impact evaluation methods for Act 129 programs have evolved over the course of the Pennsylvania Act 129 programs. Savings from established program offerings that involve proven and well-tested technologies can generally be calculated using algorithms listed in the applicable Pennsylvania TRM section or IMPs. Basic levels of rigor are typically applied when verifying established measures. EDC evaluators conduct inspections, surveys, or desk audits of a random sample of installations to determine if measures are installed and operating. Verified gross program savings are then calculated based upon the findings of the verification activity.

Nascent technologies and innovative program designs often require more in-depth impact evaluation work because the energy and peak demand savings are not as well-understood in the industry. It is the EC’s responsibility to help the EDCs identify what works and what does not so that the EDCs can scale promising offerings and minimize investment in offerings that do not deliver benefits.

According to the hierarchy within the process of implementing and evaluating EDC programs, the TRM savings protocols for efficiency measures define how ICSPs generally will calculate the ex ante savings. The impact evaluation protocols are the procedures the EDC evaluators must follow to verify the energy and demand savings claimed by the ICSPs, as defined in this Evaluation Framework. Open communication between ICSPs and ECs helps reduce or eliminate redundant data collection efforts when appropriate. However, ICSPs cannot estimate verified gross savings. TRM protocols (Section 2.3.3) have evolved over the course of Act 129 implementation and should be consistently followed by

ICSPs and EDC evaluators to improve the correlation of ex ante and ex post savings. Gross verified savings estimation for certain offerings that are not well-suited to TRM measure characterizations should follow the Measure-Specific Evaluation Protocols (MEPs) in Section 6 of this framework.

3.3.2 Calculating Verified Gross Savings

The primary research objective of Act 129 impact evaluation is to calculate gross verified savings, which are the savings achieved by the program as calculated by an independent third-party evaluator. ECs should produce an independent estimate of the program and portfolio energy and demand impacts according to the appropriate savings protocols described in the SWE-approved EM&V plan. In most cases, the evaluator and ICSP will use the same savings protocol, so the evaluator's duties may be characterized as verification. For sampled projects, evaluators should verify that the key parameters and open variables the ICSP used to estimate savings align with the installed equipment at the participating home or business. If needed, evaluators should conduct independent end-use level measurements to determine the ex post values for high-impact and high-uncertainty parameters. Higher levels of rigor are particularly important for projects with combined measure savings above the TRM thresholds. For program evaluations that rely on sampling, these independent estimates should be compared to the claimed savings for a sample of sites within each program to calculate a realization rate. This realization rate should then be applied to the population of participants to determine the verified gross savings. When appropriate, the collective results of these EDC impact evaluations will also be used to inform updates to the TRM protocols so that the TRM reflects the latest available information on measure and program savings. The following subsections provide detailed guidance for EDC evaluators for calculating verified gross savings for impact evaluations.

3.3.2.1 Measure Type

Most of the savings claimed by the Act 129 programs should be reported and verified through methods described in the TRM. As noted in Section 2.3.3, each of the three measure categories (deemed, partially deemed, and custom) dictate use of specific M&V activities. Additionally, the approach to verifying savings should be clear, technically sound, and based on accepted industry standards. The quantification of savings is both an art and a science, as energy savings are the difference between energy that would have been used without the measure and energy that was used. In practice, engineering, empirical science, and reasonable assumptions need to be used to estimate what "would have been used" because this value cannot be measured.

TRM-based savings are either (1) deemed based on the number and characteristics of the units installed, sold, or given away; or (2) partially deemed and calculated using TRM algorithms in combination with open or default parameter values. Some measures in the TRM can be deemed or partially deemed depending on whether the "EDC Data Gathering" option is used. Consider measure 2.4.8 (ENERGY STAR Clothes Washers) in the 2026 TRM. Certain parameters, such as the number of cycles per year, are stipulated, meaning EDCs must use the TRM value. Other parameters, such as the share of water heaters that are electric (versus fossil fuel), have both a default value and an option for EDC Data Gathering. If an EDC wishes to pursue the deemed pathway, they can simply look up the per-unit kWh and kW savings based on the loading type (top-loading or front-loading) and capacity range.

These deemed savings were computed using the stipulated or default values for all parameters. EDCs could also follow a partially deemed approach to this measure and use unit-specific capacity and efficiency values in a line-by-line algorithmic calculation. It is common for ICSPs to rely on a deemed approach for simplicity, while the evaluator uses a partially deemed approach to calculate verified gross savings for this type of measure.

The TRM does not provide default values for all parameters. For some measures, "EDC Data Gathering" is required. Figure 4 shows an example from measure 2.2.1 of the 2026 TRM. ICSPs must collect the rated capacity of each program-supported unit and use that value to determine the reported savings. Evaluators must then independently confirm the capacity of each unit in their sample based on manufacturer specifications, AHRI lookups, or other means.

Figure 4: Example of Required EDC Data Gathering

DEFINITION OF TERMS

Table 2-9: Terms, Values, and References for High Efficiency Equipment: ASHP, CAC, GSHP, PTAC, PTHP

Term	Unit	Value	Sources
$CAPY_{cool}$, The cooling capacity of the equipment being installed	$kBTU/hr$	EDC Data Gathering	AEPS Application; EDC Data Gathering
$CAPY_{heat}$, The heating capacity of the heat pump being installed (Auxiliary electric resistance heat is not included)	$kBTU/hr$	EDC Data Gathering	AEPS Application; EDC Data Gathering
$CAPY_{RAC}$, The cooling capacity of the room AC for the RAC cooling baseline	$kBTU/hr$	EDC Data Gathering	EDC Data Gathering
$kW_{spaceheat}$, The heating capacity of the space heaters in kilowatts.	kW	EDC Data Gathering	EDC Data Gathering

For certain measure categories, the PUC publishes Microsoft Excel calculators which implement the TRM calculations in a standardized and transparent manner. Metering, building energy simulations, or other project-specific data collection activities may be required for specific parameters of partially deemed measures depending on project size.

3.3.2.2 Level of Engineering Rigor

The level of engineering rigor is defined as the level of detail involved in the verification of the EDC-reported impacts and defines the minimum allowable methods to be used by the EDC ECs to calculate ex post savings (verified gross savings). This Evaluation Framework establishes a minimum level of detail to ensure that the verified gross savings are at the level of accuracy needed to support the overall reliability of the savings in reference to statutory savings targets. The Framework also provides guidelines on the evaluation methods the ECs must use for specific evaluation groups. These groupings consist of multiple programs (program components/measures) having common characteristics that provide evaluation efficiencies in the contracting, supervision, and implementation of evaluation efforts.

The Evaluation Framework defines two levels of rigor: basic and enhanced. Each level of rigor provides a class of minimum allowable EM&V methods, based on standard evaluation practices, to offer flexibility for the ECs to assess and propose the most accurate and cost-effective methods to verify gross savings while balancing cost and rigor. The choice of basic rigor versus enhanced rigor will depend on the type of measure; relative complexity of savings calculations; level of uncertainty; and, most importantly, savings impact. Generally, ECs are allowed to choose the appropriate level of rigor, if they follow the guidelines in this section and the TRM, including the exceptions listed by impact stratum shown in Table 15. Further, the SWE reserves the right to challenge the level of rigor planned by the ECs and request revision of the verification technique prior to the evaluators' primary data collection efforts, if necessary. After primary data collection is complete, the SWE may recommend revisions to the level of rigor or verification technique to be used on similar future sampled sites.

Table 14 provides guidelines regarding the minimum allowable methods associated with the two levels of rigor. Evaluators are highly encouraged to collect additional data that may be useful for determining the necessity of future TRM updates that improve the accuracy and reliability of savings protocols.

The EM&V options defined under each level of rigor provide ECs cost-effective methods to verify program impacts without compromising the accuracy of the reviews. In general, the TRM fully deemed measures would follow a basic level of rigor, while custom measures will typically follow an enhanced level of rigor.[31] Partially deemed measures will follow either a basic or an enhanced level of rigor, depending on the type of measure, exceptions noted by impact stratum, and level of impact. Certain measures, like behavior modification, will require a specific protocol defined in the Evaluation Framework (Section 6.1). These paths are depicted in Figure 5, which provides guidance on choosing the level of rigor by measure type.

[31] Low-impact and low-uncertainty custom measures may use a basic level of rigor.

Figure 5: Expected Protocols for Impact Evaluations

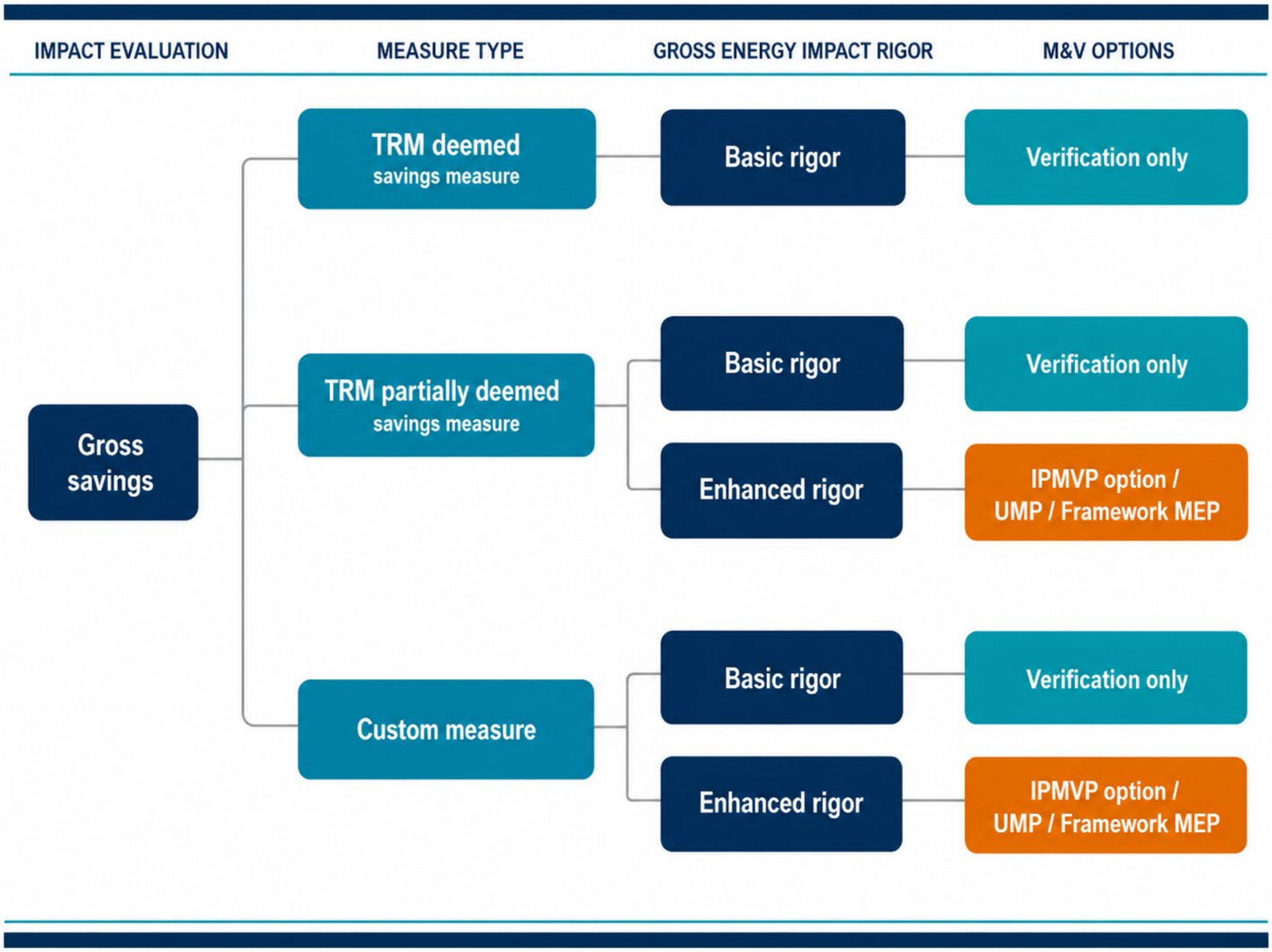


Table 14: Required Protocols for Impact Evaluations

Rigor Level	Minimum Allowable Methods for Gross Impact Evaluation
Basic	Verification-only analysis for TRM fully or partially deemed measures with impacts below the threshold established in the TRM for requiring customer-specific data collection. Verification of the number of installations and the selection of the proper deemed savings value from the TRM. Verification of appropriate application of the TRM savings algorithms for TRM partially deemed measures using site-specific data that is typically limited to equipment capacity, efficiency, and configuration. Verification of appropriate application of the savings algorithms for low-impact measures using site data that is typically limited to equipment characteristics and does not need to be measured on-site.
Enhanced	Engineering model with EM&V equal to IPMVP Option A for TRM partially deemed measures. Required for impacts above the threshold in the TRM. When the TRM specifies an algorithm, this approach includes verification of the appropriate application of TRM savings algorithms and corresponding site-specific stipulations as required and allowed by the TRM. Spot measurement and site-specific information can be obtained by the implementer and verified by the EC or obtained by the EC directly. Retrofit Isolation Engineering methods, as described in IPMVP Option B.

Rigor Level	Minimum Allowable Methods for Gross Impact Evaluation
	A regression analysis (IPMVP Option C)[32] of consumption information from utility revenue meter with adjustments for weather and overall period reported. The SWE Team recommends that at least twelve months of pre- and post-retrofit consumption be used when practicable, unless the program design does not allow for pre-retrofit billing data, such as new residential construction. In these cases, well-matched control groups and post-retrofit consumption analysis are allowable. Building energy simulation models as described in IPMVP Option D.

For partially deemed measures that require project-specific data collection and custom measures, it is recommended that the ICSP follow a similar approach to collect this information during application processing or the rebate approval process. The impact assessment methodologies used by the ICSPs and ECs should ideally be aligned to increase the correlation of ex ante and ex post savings estimates to improve the precision of evaluation results. ECs can leverage information collected by the program ICSPs in cases where it would be burdensome to the participant for the EC to gather information, such as end-use metering, independently. Evaluators should exercise their professional judgment in testing the credibility and validity of the measurements gathered by ICSPs. The SWE reserves the right to challenge the evaluators' assessment of the ICSP data and may conduct independent measurements for any project in the population.

The following section provides additional detail on the basic and enhanced levels of engineering rigor to assess ex post savings for energy and demand impacts.

Basic Rigor Option 1: Verification-Only Analysis

The first class of allowable methods for basic rigor is a verification-only analysis. This analysis applies mainly to the TRM fully deemed measures, but also may be used for TRM partially deemed measures with impacts that have low uncertainty and are below the threshold established in the TRM for requiring customer-specific data collection. The objective is to confirm that measures are installed and operational, and that the installation meets required standards. Installation verification should be conducted for a random sample of projects claiming energy savings. Verification may be completed by using one of the following methods: in person, over the phone, via a web survey, or through virtual inspections. In some cases, verification can be conducted through a review of project documentation if the project documentation includes invoices, photos, or other files that allow the evaluator to conclusively determine that the measure was installed at the location. For each program, EDC evaluation plans should specify whether on-site inspections are planned, and if so, whether ECs or implementation contractors will conduct these inspections. Sampling of measures within a project and sampling at the program level for evaluation purposes should be specified according to the Sampling and Uncertainty Protocols described in Section 3.6.3.

[32] Further information on statistical billing analysis is available in *Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*, Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol. Weblink

EDC evaluation and sampling plans for Midstream programs shall address the verification approach tailored for the energy efficiency measure, TRM section requirements, program design, and participant data collected by ICSP.

Energy efficiency kits require special attention because installation rates have been found to be relatively low. EDC ECs should independently verify the installation rate of kit measures by sampling kit participants. Stratification by measure, kit type, and customer type is encouraged (see Evaluation Precision Requirements Protocol of Section 3.6). Samples should be sufficient in size to capture installation rates for kit measures that could be relatively low. Surveys should be analyzed to verify the quantity, efficiency level, and qualification of the installed measure. EDCs may choose to distribute a survey with the kits to facilitate data collection. Evaluators should leverage these survey efforts to collect data on as many other open variables as possible. For example, a survey of kit recipients who received low-flow aerators or showerheads should ask respondents to provide their domestic hot water heating fuel and these responses should be used to estimate gross verified savings.

The basic rigor level for the gross demand impact protocol prescribes that, at a minimum, summer and winter peak demand savings be estimated using coincidence factors or energy-to-demand factors defined in the TRM.

Basic Rigor Option 2: Engineering Model Without Measurement

The second class of allowable methods for basic rigor is a verification of the appropriate application of the TRM savings algorithms using documented equipment and site data without on-site measurement. If the ICSP uses the "EDC Data Gathering" option for one or more parameters, ECs should attempt to confirm the accuracy and appropriateness of the values. EDC evaluation and sampling plans for Midstream programs shall address the verification approach for energy efficiency each unique measure based on the level of participant data collected by ICSP.

The basic rigor level for the gross demand impact protocol prescribes that, at a minimum, summer and winter peak demand savings be estimated using coincidence factors or energy-to-demand factors defined in the TRM.

Enhanced Rigor Option 1: Engineering Model with Measurement

The first class of allowable methods for enhanced rigor is an engineering model with measurement of key parameters. This approach is equivalent to IPMVP Option A. The IPMVP provides overall guidelines on M&V methods; however, more program- or technology-specific guidelines are required for the EDC programs. Engineering models are straightforward algorithms for calculating energy impacts for measures such as high-efficiency lighting, appliances, motors, and cooking equipment (partially deemed measures). Several algorithms have open variables and require additional site-specific data or measurements. The TRM measure attributes that encourage project-specific data collection will be identified by providing the option of "EDC Data Gathering" in addition to a default value. Direct measurements could be from the utility revenue meter, an EMS system, or field measurement. If the methodology and data used can readily provide an 8,760 savings profile, one should be calculated for the project.

The enhanced rigor level for the gross demand impact protocol prescribes that, at a minimum, summer and winter peak demand savings be estimated based on the allocation of gross energy savings using coincidence factors, load shapes, or energy-to-demand factors from the TRM. While direct measurements are required for this option, the measurements do not need to be performed during both the summer and winter peak demand periods. Based on the timing of project completion and evaluation, evaluators will often need to take measurements in one season and annualize the energy and peak demand savings based on engineering judgment and available information. Alternatively, peak demand savings may be informed by load shapes derived from comprehensive, statewide residential and commercial lighting studies[33] and other similar statewide or regional load shape studies.

Enhanced Rigor Option 2: Retrofit Isolation Engineering Models

The second class of allowable methods for enhanced rigor is the retrofit isolation measurements, as described in Option B of the IPMVP. This method is used in cases where full field measurement of all energy use parameters for the system in which the efficiency measure was installed is feasible and can provide the most reliable results in an efficient and cost-effective evaluation. One typical example where such a method would be appropriate is a large solar photovoltaic project where the electrical output of the array is directly measured at the inverter.

The enhanced rigor level for the gross demand impact protocol prescribes that, at a minimum, summer and winter peak demand savings be estimated based on the allocation of gross energy savings using coincidence factors, load shapes, or energy-to-demand factors from the TRM. While direct measurements are required for this option, the measurements do not need to be performed during both the summer and winter peak demand periods. Based on the timing of project completion and evaluation, evaluators will often need to take measurements in one season and annualize the energy and peak demand savings based on engineering judgment and available information.

Enhanced Rigor Option 3: Billing Regression Analysis

The third class of allowable methods for enhanced rigor is a regression analysis of consumption data that statistically adjusts for key variables that change over time and are potentially correlated with consumption. As a way of capturing the influence of weather, evaluators may incorporate weather-normalized consumption as the dependent variable or include heating- and cooling-degree days, or another explanatory variable describing the weather, directly in the model. Other variables that often are correlated with consumption include manufacturing production, the state of the economy (recession, recovery, economic growth), fuel prices, occupancy changes, behavior changes (set-points, schedules, frequency of use), changes in operation, and changes in schedule. ECs are free to select the most appropriate additional variables to include. In certain cases, selecting matching control groups

[33] See *Act 129 SWE Commercial and Residential Light Metering Study:* Weblink

may be required to calculate differences between the treatment (participant) and control groups' pre- and post-consumption. A comparison group approach is beneficial to isolate non-programmatic, extraneous effects and determine the true impact of the program intervention.

This option is the foundation of meter-based savings programs such as Retro-commissioning (RCx), Virtual Commissioning (VCx), Strategic Energy Management (SEM), or Pay-for-Performance. These programs are sometimes referred to as "M&V 2.0," Advanced M&V, or Normalized Metered Energy Consumption (NMEC). EDCs and their ICSPs may have any number of business arrangements with respect to estimating project performance, compensating program participants or contractors, and claiming reported gross savings. However, gross verified savings for projects that rely on a pre-post analysis of utility meter data should only claim gross verified savings once a full 365 days of post-installation (or performance) data is available for analysis by the EM&V contractor unless the SWE approves an abridged performance period for the project. A full year of pre-data for model training is also recommended, but somewhat out of the control of the evaluator. This rule ensures that project performance is observed over a full range of seasonal and weather conditions. Although not required for short EUL measures, EDC EM&V contractors typically weather-normalize gross verified savings estimates for reporting. Without a full year of performance and weather data, there is potential to introduce bias via the weather-normalization procedure due to out-of-sample estimation.

If EDCs claim reported gross savings from projects without a full year of performance data, evaluators should treat the MWh and MW savings from those projects as "unverified" in the Annual Report from that program year. In the subsequent program year, once a full year of performance data is available, the gross verified savings can then be claimed. This creates a timing constraint for the last year of a phase, so we recommend EDCs communicate a cutoff date clearly to their ICSPs and communicate proactively with the SWE if project-specific flexibility is needed.

Site-specific regression models can produce noisy results, especially when the effect size is a small percentage of facility load, or the variation in facility load is not well-explained by weather, production, or other available independent variables. Statistical significance is not required at the site level since EDC programs have the benefit of aggregation across customers. ICSP model specifications for meter-based projects should be established at the time of project initiation and not revised based on the observed trends in performance period data. ECs should discuss their criteria for modifying or updating the model specifications for verified savings calculations in their Phase V EM&V plans. If a meter-based regression analysis fails to produce defensible energy and peak demand savings estimates, evaluators may select a different enhanced rigor option for evaluation. The SWE should be consulted before transitioning a project from enhanced rigor to basic rigor and allowed to review the proposed engineering calculations. Potential reasons for abandoning a meter-based approach include:

- Non-routine events in the facility that cannot be separated from the impact of the Act 129 upgrades.
- Percent savings that are implausibly high or low based on the nature of the implemented upgrades.
- The estimated savings vary widely depending on the model specification.

- The in-sample CVRMSE of the baseline or performance period regression are very poor (e.g. greater than 0.75)

For residential programs, a whole-house billing analysis is advisable for installation of measures that yield greater savings (e.g., heating and cooling equipment or insulation) or when multiple types of measures are installed in a home (for the purposes of determining the appropriateness of whole-house billing analysis, we consider an energy efficiency kit to be a single measure). These EM&V guidelines are based on the UMP Protocols, which are consistent with the IPMVP Option C (Whole Facility) for annual energy savings and coincident peak demand savings, respectively.[34] The UMP recommends utilizing a billing analysis to estimate total savings when multiple measures and retrofits have been installed on site to capture the combined effects of the installed measures or when the measure is anticipated to yield substantial savings.

The enhanced rigor level for the gross demand impact protocol requires primary data from the program participants. This data could be interval-metered data from the utility revenue meter. Connected device telemetry is typically not applicable because it is rarely available for the pre-installation period. If the methodology and data used can readily provide an 8,760 savings profile, one should be calculated for the project. Data should be used to construct pre- and post-retrofit peak-hour load shapes. The data should be adjusted for weather, day type, and other pertinent variables. In cases where regression analysis is based on monthly data, coincidence factors or energy-to-demand factors may be informed by load shapes derived from comprehensive 8,760-hourly studies utilizing customer data at the same whole building level as the energy savings analysis.

Enhanced Rigor Option 4: Whole Building Simulation

The fourth class of allowable methods for enhanced rigor is building energy simulation programs calibrated as described in the Option D requirements in the IPMVP. The engineering models that meet the Option D requirements are building energy simulation models. This method can be applicable to many types of programs that influence commercial, institutional, residential, and other buildings where the measures affect the HVAC end use. This method is often used for new construction programs and building HVAC or shell upgrades in commercial and residential programs.

In addition, industrial projects can include changes in process operations where the appropriate type of model could be a process-engineering model. These are specialized engineering models and may require specific software to conduct an engineering analysis for industry-specific industrial processes. Where these types of models are more appropriate, the gross energy impact protocol allows for the use of a process engineering model with calibration as described in the IPMVP protocols to meet the enhanced rigor level.

---

[34] *International Performance Measurement & Verification Protocol (IPMVP); Concepts and Options for Determining Energy and Water Savings: Volume 1.* Prepared by Efficiency Valuation Organization. September 2009. EVO 10000 – 1:2009. and Uniform Methods Protocols: Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol. Weblink

The enhanced rigor level for the gross demand impact protocol requires an 8,760 load profile derived from a customer specific calibrated engineering model, where the modeling approach meet all the requirements in the IPMVP protocol.

3.3.2.3 Level of Engineering Rigor Mapped to Program Stratification

The impact evaluation sample should be stratified based on the constituent projects' level of impact. The stratification method in this Evaluation Framework assumes three strata in programs with a large variety of rebated measures and associated variability of savings and potential impact. However, the stratification plan and level of rigor to be used in an evaluation will be determined and documented by the EC. The actual number of strata used will be at the EC's discretion and thus this section should be interpreted accordingly. Typically, Stratum 1 will include the projects with the highest impact and/or uncertainty measures, the lowest sampling weight, and enhanced levels of rigor. Conversely, Stratum 3 includes the projects with the lowest impact and/or uncertainty measures, the highest sampling weight, and the least rigorous evaluation expectations. Non-residential projects above the TRM thresholds should be evaluated at enhanced levels of rigor. Measures that fall into Stratum 2 require either basic or enhanced levels of rigor. If a specific measure meets one of the exceptions listed in Stratum 2 (shown in Table 15), an enhanced level of rigor is required. However, sound engineering judgment is necessary to determine the applicability of the exceptions to individual measures. Generally, flexibility is allowed in determining if these conditions are met; however, the SWE reserves the right to challenge the level of rigor used by the ECs and request revision of the verification technique for future evaluation plans. As a general guidance, complex residential offerings, such as whole-building and comprehensive measure programs, and non-residential samples below the TRM thresholds should have a 50/50 mix of basic and enhanced levels of rigor. Further, evaluators are encouraged to stratify whole-building and comprehensive measure programs by housing type (i.e., single-family and multifamily homes). Evaluators should explain the sampling plan and levels of rigor in each stratum in the annual EM&V plan.

Table 15: Definitions of Program Strata and Their Associated Levels of Rigor for Impact Evaluation of Non-Residential Programs

Stratum Level	Minimum Allowable Methods for Gross Impact Evaluation
Stratum 1 – High-Impact and/or High-Uncertainty Measures	Enhanced rigor. Projects above the TRM thresholds should be in this stratum
Stratum 2 – Medium-Impact and/or High-Uncertainty Measures	Either an enhanced or a basic level of rigor may be used, depending on the applicability of the exceptions listed in this table cell and the VOI. As a guide, enhanced rigor should be used if the measure meets one or more of the following criteria: 1. Irregularity of loads: there is no sufficient pattern to predict loads with ease and accuracy 2. Irregularity of operating periods: there is no sufficient pattern to predict operating periods with ease and accuracy 3. Savings consistency: a one-time snapshot assessment likely does not capture the savings over time (e.g., measures heavily dependent upon human interaction/control)

Stratum Level	Minimum Allowable Methods for Gross Impact Evaluation
	4. High probability of substantial variance in savings calculated from a default value in the TRM 5. Significant interactive effects like whole building programs, which are not already considered in the TRM, exist between measures. An interactive effect is considered significant if the EDC EC suspects that inclusion of interactive effects in the impact estimates for the project has the potential to increase or decrease the energy or demand savings by more than 15% The projects in this stratum should strive for a 50/50 mix of basic and enhanced levels of rigor.
Stratum 3 – Low-Impact Measures	Basic rigor. Custom projects may be in this stratum if they meet both of the following criteria: 1. Are less than 50,000 kWh in energy savings 2. Utilize a reliable partially deemed savings algorithm from an established industry source, such as ENERGY STAR or a non-PA TRM

*The EDC and EC may determine the appropriate level of impact and uncertainty when stratifying measures. The EDC and EC's discretion also includes determining the relative impact of programs within the portfolio when determining level of rigor to be used. For example, the "high- impact/uncertainty" stratum of a program with relatively lower savings may not require as rigorous evaluation activities as the "high-impact/uncertainty" stratum of a program with relatively much larger savings.

3.3.3 EM&V Activities

This section provides a list of EM&V methods that are acceptable for verified savings estimation, separated per the level of engineering rigor discussed in Section 3.3.2.2.

3.3.3.1 Changes in Measure Installation Performance

In certain conditions, the EC may find that a measure was uninstalled or not currently operating, but the ICSP reported that the measure was installed and correctly operating. For example, if the measure was removed or no longer operating because of a broader change in the home or business, such as a firm going out of business, but the ICSP reported that the measure had been installed and correctly operating, the EDC can claim the savings as verified. In these conditions, appropriate savings may be considered verified and shall be calculated using default TRM parameters for the customer site. In cases where there is no applicable default TRM parameter, ECs should review the ex ante savings methodology and assumptions for accuracy in representing the measure's originally intended operation and adjust the assumptions and/or methodology as needed to determine verified savings.

This approach is permissible because it is understood that the measure's effective useful life is a market average, and these values would include conditions of premature removal. However, if the measure was removed or replaced by the participant due to dissatisfaction with the program-supported equipment, the EDC cannot claim the savings as verified. The SWE reserves the right to reconsider this assumption on a case-by-case basis.

This allowance does not apply to measures where the ICSP does not confirm installation, such as a giveaway for a direct mail kit. Under these conditions, the verified savings shall include an in-service rate (ISR), where defined in the TRM, to address uninstalled applications of measures.[35]

3.3.3.2 Basic Rigor EM&V Activities

Baseline Assessment

At a basic level of rigor, both early replacement and replace-on-burnout scenarios leverage TRM assumptions regarding the baseline equipment case. The EC should verify that TRM assumptions are appropriate for the measure delivery option being evaluated.

Measure Installation Verification

The objectives of measure installation verification are to confirm that the measure was installed and operating correctly, the installation meets reasonable quality standards, and the measure is expected to continue to generate the predicted savings during its effective useful life. At a basic level of rigor, phone interviews and/or web surveys, combined with appropriate invoices and manufacturer specification sheets, may be used to verify the measure type.

If the EC finds that a measure is operating, but in a manner that renders the TRM values not directly applicable, TRM deemed values should not be directly applied and the EC must incorporate the noted differences in savings calculations. When possible, measure design intent (i.e., the designed measure function and use and its corresponding savings) should be established from program records and/or construction documents. If the TRM values were applied incorrectly, the evaluator should recalculate savings using the correct TRM values applicable to the measure.

3.3.3.3 Enhanced Rigor EM&V Activities

Baseline Assessment

Where applicable and appropriate, the SWE recommends that ECs conduct pre-installation inspections to verify the existing equipment and gather the equipment baseline data to compute the partially deemed or custom savings estimates. The first objective is to verify that the existing equipment is applicable to the program under which it is being replaced. Additionally, the baseline equipment energy consumption and run-time patterns may be established to complete the engineering calculations used to estimate savings. At an enhanced level of rigor, early replacement of existing equipment values should be verified by on-site inspection when possible and replace-on-burnout baseline equipment values should be based on local or federal minimum codes and standards. If it is not possible or feasible

[35] For incented measures that have been installed but are not being used because there is no occupant or will not be used until another, unrelated installation/project is completed, the in-service date (ISD) will be considered the date at which the equipment is energized. See section 1.3 of Volume 1 of the 2026 TRM. Weblink

for the EC to directly verify early replacement of functional existing equipment, ECs should work with ICSPs to obtain documentary evidence of the existing equipment and its operation.

Measure Installation Verification

Evaluation plans should describe site inspections planned for residential and non-residential programs. At an enhanced level of rigor, measure installation should be verified through on-site inspections of homes or facilities. Equipment nameplate information should be collected and compared to participant program records as applicable. Sampling may be employed at large facilities with numerous measure installations. As-built construction documents may be used to verify measures, such as wall insulation, where access is difficult or impossible. Spot measurements may be used to supplement visual inspections, such as solar transmission measurements and low-e coating detection instruments, to verify the optical properties of windows and glazing systems.

Correct measure application and measure operation should be observed and compared to project design intent. For example, for C&I, ECs should note LED applications in seldom-used areas or occupancy sensors in spaces with frequent occupancy during measure verification activities then modify HOU categories appropriately. Further, if the EC finds that a measure is not operating in the manner specified in the TRM, they should not apply the TRM deemed values directly, and they must incorporate the noted differences in savings calculations. For example, if the EC discovers that a chiller is being used in an application other than comfort cooling, they should not use the TRM algorithm based on comfort cooling operating characteristics. In addition, they should obtain and review commissioning reports (as applicable) to verify proper operation of installed systems. If measures have not been commissioned, measure design intent should be established from program records and/or construction documents. Functional performance testing should be conducted, when applicable, to verify equipment operation in accordance with design intent.

On-Site Sampling of Installations

This section provides guidance in determining the number of installations to verify during the on-site inspection of a large project, such as a lighting retrofit with several thousand fixtures within a facility. The methods explained below are not exhaustive, and ECs are encouraged to propose other options in their program evaluation plans.

The first method is to verify a census of all of the installations on-site. This activity is to be done in cases where a limited number of installations were made, or when there is a large variance in operating parameters across high impact installations. For projects where a visual inspection of each installed measure would require excessive time or facility access, a statistically valid sample can be used. Samples of measures selected for verification at a particular site should be representative of all measures at the site and should be selected at random. Measures within a site should be grouped according to similar usage patterns, thus reducing the expected variability within each usage group. Within each usage group, the sampling unit should be the individual measure, with the goal being to verify the measure quantity recorded in the program tracking data.

When verifying installation quantities, the recommended relative precision for sampling on-site installations is ± 20% at the 90% confidence level at the facility level. The sampling unit (line item on the TRM Appendix C form,[36] condensing unit, appliance, etc.) should be identified in the SSMVP for custom measures. The initial verification proportion (p) assumption for determining the minimum sample size for binary (fully deemed) outcomes should be set at 50% as this will maximize p*(1–p) and guarantee that precision targets are met. For continuous outcomes, such as the number of fixtures within a space on the TRM Appendix C form, a Cv of 0.5 is appropriate.

The sample, in general, should be representative of the population; this is where stratification will be of great use. Measures with similar operating characteristics and end-use patterns should be grouped into homogeneous strata and the sampling algorithm should be designed to achieve 90/20 confidence/precision for each site. For example, lighting retrofits in common areas should be separated from those in individual suites in an office building, or air handler unit (such as a fan) motor retrofits should be grouped separately from chilled water pump replacements for C&I applications.

Since a certain degree of uncertainty is expected with any on-site counting exercise, an error band[37] should be specified within which the claimed installations or savings will be accepted. The SWE recommends using a maximum 5% error band. The error band should be calculated based on the sampling unit. If the verification counts for each usage group in the sample are within +/- 5% of the reported counts, the installed quantity should be accepted at the claimed value. For example, if the program tracking record for a project claims that 240 fixtures were retrofitted in the hallways of an office building, but the EC only counts 238 fixtures, it is not necessary to adjust the claimed fixture count in the ex post savings calculation (because the error is within +/- 5%). However, if the EC verifies only 210 fixtures in the facility hallways, ex post savings values should be calculated based on the evaluator's observations.

Site-Specific Measurement and Verification Plan

A SSMVP is designed to specify the data types and collection techniques necessary to verify savings. SSMVPs for projects within a prescriptive program will be very similar. A common plan is typically updated with the specifics of each project prior to the site visit. For custom measures, SSMVPs are individually created for each project in the evaluation sample. The ECs must draft a SSMVP for each project, which will include all measures and define the quantitative data that must be collected from the field with the corresponding source(s) for that data. SSMVPs are required for projects with combined measure savings above the TRM thresholds and are encouraged for all projects. The SSMVP should cover all activities dedicated to collecting site-specific information and any assumptions necessary to calculate savings according to the methodology specified in the SSMVP. For non-custom measures, general measure-specific data collection workbooks may be used for preparing and completing on-site visits. For custom measures, the SSMVP should include a full narrative describing all

[36] TRM Appendix C. Weblink
[37] This error band is applied solely when verifying the ex ante savings (that is, when calculating the ex post savings and determining the realization rate).

of the associated evaluation activities and ensuing calculations. These activities typically include the following:

- Measure counts
- Observations of field conditions
- Building occupant or operator interviews
- Measurements of parameters
- Metering and monitoring

For custom measures, special considerations should be taken into account for developing SSMVPs. Field measurements are an important component of determining savings for complex projects. The SSMVPs should follow the requirements of the IPMVP. Note that the IPMVP is written to allow for flexibility, but its application requires a thorough knowledge of measure performance characteristics and data acquisition techniques. Energy use varies widely based on the facility type and the electrical and mechanical infrastructure in the facility or system. A measurement strategy that is simple and inexpensive in one building (such as measuring lighting energy at a main panel) may be much more expensive in a similar building that is wired differently. For this reason, evaluation resources, costs, and benefits must be considered and allocated given the type of measure and its impact.

ECs should assess the expected uncertainty in specific engineering parameters and develop a SSMVP that manages the uncertainty in the most cost-effective manner.

The SSMVP for sampled measures should include the following sections:

1. Goals and Objectives
2. Building Characteristics and Measure Description
3. EM&V Method
4. Data Analysis Procedures and Algorithms
5. Field Monitoring Data Points
6. Data Product Accuracy
7. Verification and Quality Assurance Procedures

The content of each of these sections is described below.

Goals and Objectives: The SSMVP should state explicit goals and objectives of the EM&V.

Site Characteristics: Site characteristics should be documented in the plan to help future users of the data understand the context of the monitored data. The site parameters to be documented will vary by program and measure. The site characteristics description should include the following, as relevant:

- Site use type, such as industrial manufacturing, hotel, apartment building, retail store;
- Building configuration and envelope characteristics, such as building floor area, conditioned floor area, number of building floors;

- Building occupant information, such as number of occupants, occupancy schedule, and building activities;
- Internal loads, such as lighting power density, appliances, and plug and process loads;
- Type, quantity, capacity, and nominal efficiency of relevant heating and cooling systems;
- System control set points or sequences of operation;
- Changes and non-routine events occurring during the monitoring period that may affect results, including modifications to occupancy, equipment operation, equipment control schemes, and industrial production; and
- Description of the ECMs at the site and their respective projected savings.

The SWE recognizes that not all of these site descriptions are attainable before the site visit occurs and while drafting the SSMVP. However, evaluators should include as many attainable descriptions as feasible in the SSMVP and include any remaining descriptions in the final on-site report.

EM&V Method: The EM&V method chosen for the project should be specified. EM&V methods generally adhere to the applicable IPMVP protocol for the defined level of rigor. The ECs have considerable latitude regarding the development of an SSMVP, which may be a combination of the IPMVP options.

In certain site conditions, the EM&V inspection method may be conducted through a virtual meeting and remote data collection of the appropriate parameters. The following site conditions outline where virtual inspections may be a cost-effective and preferred alternative to on-site inspection:

- Limited number of affected spaces for lighting projects
- Limited number of pieces of affected equipment
- Likely availability of EMS trend data that can be electronically transferred
- Photos of rebated equipment have already been collected by the ICSP or from the customer
- A site contact with detailed understanding of the equipment operation is available

Data Analysis Procedures and Algorithms: Engineering equations and data points for collection should be identified in advance and referenced within the SSMVP. Engineering calculations should be based on the TRM for partially deemed measures. The equations, assumptions, and documentation supporting the defined baseline and measure savings calculations may be presented in the most convenient format (spreadsheet or written report) but should always be clearly stated and explained.

Field Monitoring Data Points: If any field measurements are planned, they should be specified, including the sensor type, location, frequency of measurements, duration of measurement period, and engineering units.

Data Product Accuracy: When field measurements are planned, the accuracy of the planned instrumentation should be included in the SSMVP. This information is presented in the specification sheet for most commercially available data logging equipment.

Where measurements may need to be normalized or annualized to another parameter, the SSMVP shall describe the normalization rationale, expected algorithm for pre- and post- conditions, and the source of the non-measured data. Rationale for normalized measurement may include, but is not limited to, correlation of weather, production, and/or occupancy changes. For example, in a situation where the ECs intend to annualize savings using a comparison of the production levels from a plant during the M&V period to an estimate of annual production of the facility, this section should discuss the source and basis for the annual production estimates.

Verification and Quality Assurance Procedures: Data analysis procedures to identify invalid data and treatment of missing data and/or outliers must be provided. This should include quality assurance procedures to verify data acquisition system accuracy and sensor placement issues. For example, to ensure the accuracy of trend data for a motor's power consumption recorded in an EMS, ECs can compare EMS trend data to what would be expected from the equipment manufacturer's performance specifications. If the EMS data do not align with what is expected, ECs should then take additional steps to verify EMS readings, such as taking a contemporaneous spot measurement with a calibrated handheld device.

3.4 Net Impact Evaluation

The PUC stipulated in the Phase V Final Implementation Order that compliance in Phase V be determined using gross verified savings and that NTG research results will be used for modifications to existing programs and for planning purposes for future phases.[38]

The PUC, however, recognizes that NTG findings and NTG-based TRC ratios provide all stakeholders with additional information regarding the effectiveness of EE&C measures and programs.[39]

EDCs' ECs should therefore conduct NTG research and consider conducting additional research to assess market conditions and market effects to determine net savings. Market effects research is discussed in Section 3.4.1.3.

When conducting NTG research, the NTG methods should be consistent across time and EDCs.[40] If the NTG metric is measured the same way across time, program staff can use the NTG metric to inform their thinking because it provides a consistent metric over time. Another reason for a uniform NTG approach is that the value that can be obtained from comparing NTG metrics across utilities. Just as programs change year to year, it is clear that the programs offered by the EDCs vary from each other. When there are different metrics, no one can discern whether different NTG values are due to program differences, external differences, or differences in the metric. By using a consistent metric, program

---

[38] *Phase V Final Implementation Order*, at page 221. Weblink
[39] *Phase V Final Implementation Order*, at page 221. Weblink
[40] However, with new programs or program delivery methods, evaluators will need to assess the most appropriate NTG methods to employ.

staff can at least rule out differences in the metric as the reason. EDCs should, however, provide both gross and net verified energy and demand savings in their final annual reports.

The SWE notes that net impact evaluations of low-income programs are not required, and the EDCs can assume a NTG ratio of 1.0 for low-income programs. Free riders are not anticipated among low-income participants due to income constraints.

3.4.1 Acceptable Approaches to Conducting NTG Research

NTG research traditionally has two primary purposes: (1) attribution (i.e., adjusting gross savings to reflect actual program influence on savings) and (2) explicating customer decision-making and the contribution the program made to the customer's decision to install an energy efficient solution. This research helps to determine whether a program should be modified, expanded, or eliminated based on its Net-to-Gross Ratio (NTGR).

The UMP provides the following relevant definitions: [41]

- **Net Savings**: Changes in energy use that are attributable to a particular EE program. These changes may implicitly or explicitly include the effects of free ridership, spillover (SO), and induced market effects.
- **Free Ridership**: Program savings attributable to free riders (program participants who would have implemented a program measure or practice in the absence of the program).
- **Spillover**: Additional reductions in energy consumption or demand that are due to program influences beyond those directly associated with program participation.
- **Market Effects**: A change in the structure of a market or the behavior of participants in a market that is reflective of an increase in the adoption of energy efficiency products, services, or practices and is causally related to market intervention(s). According to Prahl et al., "Market effects are best viewed as spillover savings that reflect significant program-induced savings in the structure and functioning of energy-efficiency markets." [42]

Program evaluators traditionally use one of several methods to assess a program's net savings, including self-report surveys, econometric methods, market sales data analysis, comparison area analysis, top-down evaluations, structured expert judgment, and historical tracing, many of which may be used to assess market effects. The UMP details these various methods.[43] Much has been written

[41] Violette, Daniel and Pamela Rathbun, "Estimating Net Savings: Common Practices," in *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures.* Prepared for the National Renewable Energy Laboratory, October 2017. Weblink

[42] Prahl, R., R. Ridge, N. Hall & W. Saxonis. 2013. "The Estimation of Spillover: EM&V's Orphan Gets a Home." In Proceedings of the 2013 International Energy Program Evaluation Conference. Chicago, August 13-15. Accessed April 27, 2026. Weblink

[43] Violette, Daniel and Pamela Rathbun, "Estimating Net Savings: Common Practices," in *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures.* Prepared for the National Renewable Energy Laboratory, October 2017. Weblink

about the various methods and their relative strengths and weaknesses.[44] In light of increasing program activity, as well as activity external to the program that contributes to customers' engagement with energy efficiency, net savings estimation is difficult to compute. The most cost-effective measurement technique for net savings is self-report surveys; however, social science research shows that measurement of the counterfactual (what would have happened in the absence of the program) using self-reports can be problematic. In addition, while increased participant and non-participant spillover installations may be making a greater contribution to savings than the amount that free ridership detracts from savings, measuring spillover using self-reporting suffers from similar problems to those stemming from using it to measure free ridership, and when on-site confirmation is included, it becomes very costly.[45]

Other methods, however, may be even more costly. In particular, with econometric and comparison area approaches it is not possible to disaggregate the effects of free ridership and spillover, and they do not directly address customer decision-making or the program's influences on decision-making. For this reason, the SWE has determined that EDCs should use survey methods for assessing free ridership and spillover for downstream programs and has provided descriptions of common methods for doing those assessments (Appendix B, Appendix C, and Appendix D). These approaches must be used for the specific programs they apply to, though they may be used in combination with other methods. The SWE has established a procedure whereby EDCs may identify downstream programs for which the common methods are not suitable; in such cases, EDCs may propose a method, subject to SWE review. In Phase V the EDCs may use methods of their own choice, including market effects approaches, to estimate NTG for midstream and upstream programs.

Section 3.4.1.5 presents an overview of common methods for assessing net impacts of midstream and upstream programs. The SWE notes that the EDC's Phase V EE&C plans include a broad range of midstream and upstream program offerings. It is important to note that midstream and upstream programs may generate market effects if they are designed to influence manufacturers, distributors, and installers who will in turn influence their customers and the overall market. While market effects can be difficult to measure because their reach goes beyond program participants, their cumulative impact on influencing the entire market may be large and sustained over time. Net impact evaluations of midstream and upstream programs that do not take market effects into account risk missing spillover savings and thus underestimating program impacts, NTG ratios and net-TRC ratios.

The primary concern of the SWE is whether the EDCs' NTG evaluations are helping the EDCs fully understand the effects/attribution of their programs on the markets in their service territory. Further,

---

[44] A general review of issues is provided in Haeri, H. and M. Sami Khawaja, "The Trouble with Freeriders," *Public Utilities Fortnightly*. March 2012. Weblink

[45] Peters, J. S. and M. McRae. "Free ridership Measurement is Out of Sync with Program Logic…or, We've Got the Structure Built, but What's Its Foundation?" In *Proceedings of the 2008 ACEEE Summer Study on Energy Efficiency in Buildings*. American Council for an Energy-Efficient Economy. Weblink

the SWE must ensure that NTGRs are reasonable and ratepayer funds appropriately support customers who need that support in order to invest in energy efficient solutions.

3.4.1.1 Using Self-Reports for Estimating Free ridership and Spillover

Using self-reports to measure free riders and spillover is subject to bias and therefore may not yield an accurate estimate of free ridership or spillover; this concern supports the PUC's decision that self-report-based NTG should not be used to calculate net savings estimates for compliance purposes.[46] However, careful application of social science methods may help mitigate biases.[47] Years of research have shown that various NTG self-report assessments tend to produce consistent results. Thus, even if they do not necessarily produce accurate estimates of net savings at any given time, they may be useful in assessing trends over time. Thus, the SWE believes that self-report assessments of free ridership and spillover may be useful in assessing changes over time or differences across programs.

- **Free ridership** – The purpose of measuring free ridership is to ensure that the program is primarily serving those who need the program to invest in energy efficiency. Over the course of many years of DSM program evaluation, evaluators have developed methods to estimate the number of free riders and then to estimate the net savings resulting only from those who required the program's support to install the energy efficient solutions.
- **Spillover** – The purpose of measuring spillover is to ensure that the program is credited with energy savings that come from participants and non-participants who install energy efficient solutions without using program resources, and do so because of the program, either as participants who take additional efficient actions (inside or participant spillover) or as non-participants who take actions the program recommends but without program support (outside or non-participant spillover).

The NTG ratio removes free ridership from the savings calculation and adds program spillover. The NTG formula is defined in Equation 1:

Equation 1: NTG Formula

$$NTG = 1 - FR + SO + ME$$

Where:

FR = *Free ridership* quantifies the percentage of savings (reduction in energy consumption or demand) from participants who would have implemented the measure in the absence of the EDC program.

SO = *Spillover quantifies* the percentage reduction in energy consumption or demand (that is, additional savings) caused by the presence of the EDC program. Spillover savings happen when

[46] *Phase IV Implementation Order*, at page 221. Weblink

[47] Haeri, H. and M. Sami Khawaja "The Trouble with Freeriders." *Public Utilities Fortnightly*. March 2012. Weblink

customers invest in additional energy efficient measures or activities without receiving a financial incentive from the program.

ME = *Market effects* savings not already captured by spillover. Some examples of these effects include increased availability of efficient technologies through retail channels, reduced prices for efficient models, build-out of efficient model lines, and an increase in the ratio of efficient to inefficient goods sold or practices undertaken in the market.

When estimating market effects and spillover independently, great care must be taken to ensure there is no double counting of spillover and market effects savings. Energy savings estimates derived through market effects methods[48] often do not differentiate the various NTG components, such as free ridership and the various forms of spillover, but rather constitute a single estimate of net savings. When this is the case, the above formula does not apply. Instead, NTG is equal to (total savings – naturally occurring savings) / within-program savings.[49],[50]

Care must be taken when developing the questions used to measure free ridership. The SWE considers the research approaches detailed in the UMP[51] as well as those used in Massachusetts[52] and those developed by the Energy Trust of Oregon[53] to constitute some of the best practices for free ridership and spillover estimation.

The UMP advises that evaluations should make special considerations for determining NTG impacts for large or complex C&I projects. For example, the UMP recommends using trained evaluation professionals rather than general telephone surveyors to address the largest, most complex projects in custom programs. For large and/or complex C&I projects, EDC evaluators should specify how NTG data collection will be conducted, by whom, and the timing of the data collection in their EM&V plans.[54] Site visits are often an ideal time to hold these technical discussions if key decision-makers from the

---

[48] For a discussion of these methods, see Rosenberg, M. and L. Hoefgen, 2009. *Market Effects and Market Transformation: Their Role in Energy Efficiency Program Design and Evaluation.* Prepared for the California Institute for Energy and Environment. Weblink

[49] NMR Group., Inc. 2014. *Methods for Measuring Market Effects of Massachusetts Energy Efficiency Programs*. Prepared for the Massachusetts Electric and Gas Program Administrators. Weblink

[50] NMR Group, Inc. 2013. A Review of Effective Practices for the Planning, Design, Implementation, and Evaluation of Market Transformation Efforts. Prepared for PG&E, SDG&E, Southern California Edison, and Southern California Gas. Weblink

[51] Violette, Daniel and Pamela Rathbun, "Estimating Net Savings: Common Practices," in *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures.* Prepared for the National Renewable Energy Laboratory, October 2017. Weblink

[52] See the most recent Massachusetts NTG methodology reports: NMR Group and Tetra Tech. 2026. Commercial & Industrial Self-report Net-to-Gross Survey and Algorithm Update. Weblink

NMR Group and Tetra Tech. 2026. Residential Self-report NTG Methodology. Weblink

For a complete list of NTG methodology reports from Massachusetts, see Appendix C.2, Footnote 112.

[53] Energy Trust Free Ridership Methodology. Weblink

[54] Violette, Daniel and Pamela Rathbun, "Estimating Net Savings: Common Practices," in *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures.* Prepared for the National Renewable Energy Laboratory, October 2017. Weblink

participating facility are present. If a customized battery of attribution questions is planned for a site-visit, the SSMVP should include details on the questions and why they are relevant for the project.

Free Rider Measurement

The SWE has determined that, where possible, EDCs should use standard sampling techniques, data collection approaches, survey questions, survey instruments, and analysis methodology for free ridership assessment. Standardization can provide consistency in explications of the programs' effects. EDCs may implement other methods concurrently.

The SWE has recommended common methodologies for estimating free ridership in downstream programs for the EDCs to use or adapt to their purposes since Phase II. One common approach applies to a broad range of incentive-based programs; the other is specific to appliance recycling programs. The SWE common approach is similar to that developed by the Energy Trust, which uses a short battery of questions but has been found to produce results that are comparable to those produced by much longer batteries.[55] The approach for appliance recycling programs is based on the approach described by the U.S. Department of Energy's UMP.

The common method uses responses to a sequence of free ridership questions to compute an overall free ridership score for each measure or program. It is very important that more than one question be used to determine the level of free ridership. Free ridership questions in the common method include two additive and equally weighted components:

- Participant intention
- Program influence

Each component provides a possible score of 0 to 50. When added, the resulting score, which has a range of possible values of 0 to 100, is interpreted as a free ridership percentage; this is also how partial free riders emerge. A score of more than 0% and less than 100% indicates a partial free rider.

Net savings for an appliance retirement program (ARP) is based on the participants' self-report of what they would have done absent the program. Savings are attributed based on three scenarios: (1) they would have kept the unit in the absence of the program but instead, as a result of the program, recycled it and did not replace it (savings equals energy usage of old unit); (2) in the absence of the program, they would have put the unit back into usage elsewhere, sold or given the unit away to another user, or sold or given away a unit that was less than ten years old to a retailer (savings equals a mix of full savings, delta old to new, and no savings); or (3) in the absence of the program, they would have taken the unit out of usage, sold or given a unit at least ten years old to a retailer, hauled it to the dump, or hired someone to discard it (free rider – no savings).

---

[55] Violette, Daniel and Pamela Rathbun, "Estimating Net Savings: Common Practices," in *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures.* Prepared for the National Renewable Energy Laboratory, October 2017. Weblink

Appendix B provides more details on the net savings approach for ARPs. Appendix C provides both the general form of questions to use and rules for calculating free ridership scores from responses to questions. As described in the Appendices, EDCs may adapt the questions to fit each program, subject to SWE review. EDCs may also add questions and/or use alternative formulas for calculating free ridership scores in parallel with the calculations resulting from the methods described in the memos.

The confidence and precision for free ridership estimates should be consistent with those for gross savings estimate requirements – that is, 85% confidence with ±15% in precision at the program level, and 90% confidence with +10% precision at the sector level. Note that this does not mean that the estimated net savings (obtained by applying the NTGR, developed from both free ridership and spillover estimates, to gross savings) must be at the 85/15 or 90/10 level of confidence/precision. Since net savings are not relevant to compliance, there is no specific precision requirement for net savings. The purpose in specifying confidence and precision levels for free ridership estimates is to ensure results that will be valuable for program planning purposes.

Spillover Measurement

Net savings claims that include spillover studies are more robust than those that include just free ridership estimates. The SWE also has determined that, where possible, EDCs should use standard techniques, instruments, and methods for spillover assessment. However, the SWE has determined that, while estimation of non-participant spillover is desirable, it is not required.

The SWE has recommended a common methodology for estimating participant and (if EDCs choose to assess it) non-participant spillover in downstream programs since Phase II. The methodology is presented in detail in Appendix D, which describes both the general form of questions to use and rules for calculating spillover scores from responses to questions. The Appendix describes the degree of latitude the EDCs have in adapting the methods. EDCs may also add questions and/or use alternative formulas for calculating spillover scores in parallel with the calculations resulting from the methods described in the memo.

The spillover approach is based on self-report. The SWE recognizes that self-reported spillover without verification may be inaccurate, and therefore the EDCs should interpret findings with caution. However, verifying spillover reports through on-site assessment is costly and therefore not required.

The common approach for participant spillover assesses for each participant:

- The number and description of non-incented energy efficiency measures implemented since program participation
- An estimate of energy savings associated with those energy efficiency measures
- The program's influence on the participant's decision to implement the identified measures
- The likelihood the participant would have implemented the identified measures without program participation.

Details of assessment and calculation of participant spillover totals and rates are provided in Appendix D.

For EDCs that choose to assess it, non-participant spillover may be assessed either through a general population (non-participant) survey or through a survey of trade allies. If a general population survey is selected, it should assess for each survey respondent:

- The number and description of non-incented energy efficiency measures implemented since program participation
- An estimate of energy savings associated with those energy efficiency measures
- The program's influence on the participant's decision to implement the identified measures.

Evaluators should submit draft survey questions to the SWE.

If an evaluator chooses to assess non-participant spillover through trade ally surveys, separate surveys should be conducted for the residential and non-residential sectors. Each survey should assess, for each sampled respondent:

- The number of program-qualified measures sold or installed within the specified sector, the specified utility's service territory, and the specified program year
- The percentage of such installations that received rebates from the specified program
- The trade ally's estimate of the proportion of their sales or installations of non-rebated measures that went to prior program participants
- The trade ally's judgment of the specified program's influence on sales of the common program-qualified but not rebated measures.

Details of assessment and calculation of non-participant spillover totals and rates are provided in Appendix D.

The SWE recommends – but does not require – that the evaluation strive to achieve confidence and precision levels sufficient to provide meaningful feedback to EDCs. If non-participant spillover is assessed, the sampling approach should produce a sample that is representative of the target population (non-participants or trade allies) or capable of producing results that can be made representative through appropriate weighting of data. In the case of trade ally surveys, the sampling plan should take trade ally size (e.g., total sales, total program savings) and type of equipment sold and installed (e.g., lighting or non-lighting) into consideration. Again, the SWE does not specify a minimum level of confidence and precision, but the evaluations should strive to achieve confidence and precision levels sufficient to provide meaningful feedback to EDCs.

3.4.1.2 Econometric Approaches

Econometric approaches may be used to estimate net savings. When used for buildings, these use historical billing data and require a non-participant group of similar buildings for which the owner has invested in end-use improvements without program support. When used for estimating changes in sales such as market lift or market share, sales data would be used.

The ideal application for econometric analysis is when customers are randomly assigned to treatment (participant) and non-treatment (non-participant) groups, such as with large-scale opt-out programs.[56] The analysis of customer billing data between the two groups distinguishes program effects and net savings. Survey data may be added to this approach to enhance the analysis and interpretation of program effects.

For opt-in or voluntary commercial-sector programs, the evaluator may conduct on-site verification of the energy efficiency level of the equipment and a survey of both participants and non-participants. A discrete choice model estimates the probability of participation, given certain characteristics and this probability is used to calculate net savings.

For opt-in or voluntary residential programs, the evaluator may use a quasi-experimental design with participants and non-participants with similar buildings. A second-stage model using survey data can facilitate inclusion of other factors, such as structural and end-user characteristics to explicate the differences between the non-participant and participant groups.

The primary disadvantages of these two approaches are (1) the difficulty in identifying comparison groups of similar buildings, or those in which new end-use equipment has been installed, and (2) the additional cost. Further, for market share approaches, it is not possible to disaggregate free riders or to identify spillover, while for matched-pair approaches, using econometric modeling provides a hybrid estimate between gross and net savings and does not provide total net savings estimates.

3.4.1.3 Market Effects Studies

Studies of market effects help estimate program effects and provide information on market needs and responses to energy efficiency programs. The purpose of measuring market effects is to make appropriate strategic decisions about program offerings and timing so that the market for energy efficient products and services may grow more readily than it would without the program.

The definition of a market effect in the California Protocols is "a change in the structure or functioning of a market or the behavior of participants in a market that result from one or more program efforts. Typically, these efforts are designed to increase the adoption of energy efficient products, services, or practices and are causally related to market interventions."[57] Only certain programs can be expected to generate substantial market effects and therefore, warrant market effects studies. Characteristics of such programs may include the following: the savings per transaction are small, but the transactions

[56] The term *opt-out* refers to a program design in which customers automatically are enrolled by the EDCs. This is common in some behavior intervention program designs where a randomly selected group of customers is provided information that other customers do not receive.
[57] TecMarket Works Team. California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals. Prepared for the California Public Utilities Commission. San Francisco, CA. April, 2006. Weblink

are numerous; the programs target markets rather than program participants; the programs aim to change energy use through changing what happens among midstream and upstream market actors, rather than focusing just on end-users of equipment or services; a significant portion of the actors in a market will be touched by the program; the programs may involve providing education or information in order to change practices or decision making that affects energy consumption; or the product or service that the program addresses offers significant non-energy benefits, such as increased comfort, increased home value, or reduced maintenance.[58]

Like the econometric models just discussed, market effects studies provide an estimate of overall market effects, from which free ridership and spillover are not disaggregated, to help in assessment of program cost-effectiveness. Failure to account for the market effects of programs that are likely to result in such effects risks undercounting net savings when assessing cost-effectiveness. Another purpose of market effects studies is to examine changes in the market and determine the source of those changes, and thus help with program design and planning. There are several factors to consider in conducting market effects studies, whenever they are appropriate based on the above criteria.[59]

1. Identify and characterize the target market (or markets) for the program.
2. Develop a theory of change against which progress is assessed. This may include a visual model or narrative describing the market and the program's interaction with it. It should also include developing metrics or market progress indicators (MPIs) against which the progress of the program in effecting change in the market may be assessed.
3. Assess progress toward the MPIs or metrics of expected change, paying particular attention to changes in market share, marketing and promotion, pricing, and product availability.
4. Market baseline measurements are very important; these form the basis of comparison and may be measure-specific or program-specific. They should be broad enough to cover possible interactions with other external influences. Baseline has two meanings in this context. For assessment of MPIs, it is a previously measured value or the starting point; for assessment of NTG, it is the counterfactual, or what would have happened in the absence of the program.
5. For assessing program cost-effectiveness, net savings attributable to market effects should be estimated.

In summary, NTGRs will not be applied when determining whether the EDCs have met their energy and demand reduction targets in Phase V of Act 129. Net savings studies such as NTG, econometric, or market effects research should be conducted for the following purposes: (1) to monitor the effects the program is having on the market, (2) to gain a more complete understanding of attribution of savings,

---

[58] NMR Group, Inc. Methods for Measuring Market Effects of Massachusetts Energy Efficiency Programs. Prepared for the Massachusetts Program Administrators and the Energy Efficiency Advisory Council. November 2014. Weblink

[59] NMR Group, Inc. 2019. Massachusetts Action Plan for Measuring Market Effects. Prepared for the Massachusetts Energy Efficiency Program Administrators. Weblink. TetraTech. 2017. Net-to-Gross Methodology Research. Weblink

(3) to identify when specific program measures no longer need ratepayer support, and (4) to help assess cost-effectiveness.

3.4.1.4 Focus on HIMs

During PY6, the SWE suggested that EDCs oversample measure categories (technologies) of high importance, called HIMs, to help program planners make decisions concerning those measures for downstream programs only.[60] The SWE proposed that for each program year,[61] each EDC identify three to five HIMs for study based on energy impact, level of uncertainty, prospective value, funding, or other parameters. The intent is to prioritize measure-level NTGRs for HIMs, but the EDCs are encouraged to also provide some program-level NTG information – that is, to over-sample HIMs, but they may also include non-HIMs in the research, as appropriate. The EDCs need not sample non-HIM measures if the HIM sample includes measures that contribute 80% of the savings to the portfolio. If an EDC evaluator believes that selection of four to five HIMs for NTGR evaluation would create an undue research burden or if it constrains the selection of non-HIM measures that may be assessed, they should indicate so in their evaluation plan and propose an approach that satisfies the intent of the requirement. The EDC evaluator's sampling plan should discuss this issue and describe its impact on non-HIM and program-level NTG assessment.

Using this method EDCs should sample HIMs at 85% confidence and 15% absolute precision to ensure the EDCs and evaluators select a large enough sample so that it is statistically valid. EDCs should combine samples for a given technology across programs or delivery channels, if it is appropriate to do so. There may be reasons why the sample should not be combined across programs or delivery channels (e.g., if it is believed that a given delivery channel or participant type may result in markedly different free ridership or spillover values than other delivery channels or participant types). The EDC evaluator's sampling plan should discuss this issue.

3.4.1.5 Approaches for Midstream and Upstream Programs

In addition to targeting consumers, upstream and midstream programs target program services and/or funding to market actors such as contractors, builders, distributors, dealers, supply houses, and manufacturers, with the goal of influencing their stocking, design, specification, recommendation, and installation practices.

---

[60] The proposed HIM-specific research does not preclude addressing custom projects at the project level only. If an EDC's evaluation contractor believes that the requirements to research and report NTGR for specific HIMs will conflict with satisfying other important NTG sampling objectives, the EDC evaluator should indicate so in its evaluator plan and propose an approach that satisfies the intent of the requirement.

[61] The proposed HIM-specific assessment does not change any prior *Framework* requirement regarding what EDC's evaluators should do if EDCs decide not to do NTG research in a given year. One suggestion, but not a requirement, is to report that no NTG research was conducted, assume the NTG is similar to prior year (that is, the same NTG ratio could be reported again), and state the reasons and rationale that were included in the evaluation plan (e.g., market conditions did not change).

In upstream and midstream programs, consumers may not be aware of program influences on sales, stocking practices, or prices. Thus, using only participant self-reports to estimate free ridership and spillover will likely result in an inaccurate estimate of net savings. In these cases, evaluators should include additional evaluation methods, such as market actor self-report surveys, to examine the effects of these upstream influences. While this leads to NTG protocols that are more involved and use multiple methods, using multiple methods allows the evaluators to triangulate and minimize the bias or error from any individual method.

There are a number of methods that are appropriate for midstream and upstream programs (particularly those with potential market effects), and evaluations may use one or more of these approaches in combination, depending on data availability and program context:[62],[63]

1. Supply-side market actor self-reported counterfactual analysis
2. Cross-sectional analysis, which may include time-series data
3. Forecasting or back casting the non-intervention baseline
4. Structured expert judgment

All these approaches, whether used individually or in combination, require each of the following:

- Estimations of the size of the market both for efficient and non-efficient measures (a.k.a. market share or market penetration) in the baseline period before the program is implemented and at the time of evaluation
- Identification of changes in market actor behavior
- Measurement of savings achieved at the market level
- Estimation of the baseline for savings (a.k.a. naturally occurring savings or the counterfactual), which is the savings that would have occurred in the absence of the program

The choice of evaluation approach will be affected by factors such as the availability of market share or market penetration data; the degree to which the market for the product, equipment, or service is already transformed; and the availability of appropriate non-program areas for comparison and degree to which they have been influenced by other areas' programs. (Market share and market penetration both refer to the ratio of sales of high-efficiency equipment to all sales of this type of equipment.)

---

[62] NMR Group, Inc. *Methods for Measuring Market Effects of Massachusetts Energy Efficiency Programs.* Prepared for the Massachusetts Program Administrators and the Energy Efficiency Advisory Council. November 2014. Weblink

[63] Violette, Daniel and Pamela Rathbun, "Estimating Net Savings: Common Practices," in *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures.* Prepared for the National Renewable Energy Laboratory, October 2017. Weblink

In addition, if the evaluators can identify a valid comparison group and control variables that influence energy use across participants and non-participants, evaluators may consider using billing data analyses with control variables and Linear Fixed Effects Regression (LFER).

Because of the widespread interest among the EDCs in midstream program offerings and because the customer base of upstream and midstream market actors may span multiple EDC territories, some of the midstream and upstream NTG protocols may be better estimated on a statewide level than at the EDC level. The EDCs should consider coordinating their NTG research, particularly for programs with similar midstream or upstream design elements and measure offerings.[64]

3.4.1.6 Accounting for Outside Funding

In the Phase V Final Implementation Order, the Commission noted the importance of the availability of funding from outside of Act 129 to help accelerate conservation programming in the Commonwealth, and the Commission stressed the importance of collaboration between the EDCs and other program administrators in its Phase V Final Implementation Order.[65] Examples of outside funding include, but are not limited to, the following:

- Inflation Reduction Act (IRA)-funded programs such as
    - The Reducing Industrial Sector Emissions in Pennsylvania (RISE PA) program
    - The Penn Energy Savers program that includes Home Efficiency Rebates (HER) and Home Electrification and Appliance Rebates (HEAR) programs managed by the Department of Environmental Protection (DEP) Energy Programs Offices
- DEP's Agricultural Energy Efficiency Rebate Program
- DEP's Solar for All Program
- Alternative Energy Portfolio Standards Act (AEPS) credits

In previous phases, guidance for measuring NTG was focused on evaluating the EDC program intervention in isolation. However, when outside funds are being leveraged, this approach can lead to underestimation of EDC attribution because customers who use public funds or resources alongside EDC offerings may not credit the EDC when asked about the EDC intervention alone. However, including public funds and resources with the EDC offerings when assessing NTG also has the potential to overestimate Act 129 attribution.

To address this potential gap in attribution, for EDC programs that are collaborating with other programs, we recommend that NTG assessments (1) ask survey respondents if their project included outside funding and (2) if so, ask respondents to attribute the impacts of the EDC program in isolation and the impacts of the combined EDC program and outside funding. This will provide the EDCs and

---

[64] EDCs with similar midstream program offerings may want to consider coordinating program implementation as well as uncoordinated midstream programs among neighboring EDCs could result in leakage, double incentives, and loss of participation.
[65] *Phase V Final Implementation Order* at 158.

stakeholders with additional information regarding the effectiveness of EE&C measures and programs as well as the EDC efforts to collaborate with outside funding.

If an EDC program has collaborated with other programs, the EDCs should report both NTG values, NTG for the EDC program in isolation and NTG for the combined impacts of the combined EDC program and outside funding. The EDC program NTG should be used for reporting NTG savings, net TRC values, etc. The combined EDC program and outside funding NTG values are for informational purposes.

3.5 Process Evaluation

The purpose of process evaluation is to determine if there are ways to alter the program to improve program cost-effectiveness or the program's efficiency in acquiring resources. Process evaluations are a significant undertaking, and they must be designed and executed systematically to ensure unbiased and useful results.

Process evaluations consist of in-depth examinations of the design, administration, delivery/implementation, and market response to energy efficiency programs. As with all evaluations, a process evaluation should address the specific program goals. While they primarily serve the EDC's program staff and management, process evaluations also provide a vehicle for sharing program design and operational improvements with other professionals in the field. Below are examples of how decision-makers can use the results of process evaluations:

- Improve program performance with respect to internal administration and communications, promotional practices, program delivery, incentive levels, and data management
- Provide a means of improving customer satisfaction and identifying market threats and opportunities
- Provide information to regulators and other interested parties that programs are being implemented effectively and modified or refined as necessary
- Provide a means of contributing to industry-wide knowledge and best practices so that other EDCs can improve their programs

This section provides a minimum set of standards for process evaluations across the EDCs' portfolios that ensure the necessary flexibility and control for program administration and management so the PUC can be confident that the EDCs manage their programs as cost-efficiently as possible.

3.5.1 Process Evaluation Approaches and Timing

Process evaluations use program data, secondary data, document review, direct observations/site visits, and a variety of one-on-one or group interviews and surveys to gather information to describe and assess programs. The design for each process evaluation should begin with the program's original design intent and should provide evidence of progress in achieving program goals and objectives from the perspective of its various target audiences. Below are examples of how decision-makers can use the results of process evaluations:

- Highlight areas of program success and challenges
- Make recommendations for program modification and improvement
- Identify best practices that can be implemented in the future

Each process evaluation should have a detailed plan that describes the objectives, sampling plan (for surveys, interviews, or focus groups), research activities, and specific issues to be addressed, along with a schedule of milestones and deliverables.[66]

Every program should have at least one process evaluation in every funding cycle or phase. The process evaluation may be either an in-depth, comprehensive process evaluation or one of several types of focused process evaluations. Process evaluations should be timed to coincide with decision points for the program design and implementation process. The primary types of process evaluations are described below:

1. Standard Comprehensive Process Evaluation – This includes data collection activities with each of the program's target audiences, including participants, non-participants, end users, and trade allies. Such complex evaluations require resources and time to implement. The New York State Process Evaluation Protocols[67] provide excellent guidance on the best practices for all process evaluations, and in-depth, comprehensive process evaluations will adhere to the majority of those protocols.
2. Market Characterization and Assessment Evaluation – Market characterization and market assessment activities are important to help program staff understand how the market is structured, operating (characterization), and responding to the program offerings (and to activities external to the program [assessment]). Such studies usually focus on specific technologies or product and service types. They are conducted in order to inform program design and redesign and may be integrated into a comprehensive process evaluation.
3. Topic-Specific Focused Evaluation – Not every process or market evaluation must be comprehensive. In cases where a comprehensive evaluation has been conducted, it may be appropriate to conduct an abbreviated process evaluation that focuses on specific items, such as program features or ideas program staff want to explore to see if changes to the program are warranted; data collection for this type of evaluation will involve targeted questions to carefully selected audiences.
4. Early Feedback Evaluations – New programs, recently updated/modified programs, and pilot programs benefit from early program evaluation feedback. Such evaluations can help program designers and managers refine the program design before full-scale rollout or during the current program cycle. These early feedback evaluations should be short and focus on as few as three to six months of program operation in order to give program staff rapid and specific feedback.
5. Real-Time Evaluation – In many cases, process and market evaluation can help programs be more effective if the information on program progress and performance can be conducted

[66] The SWE reserves the right to review the process evaluation plans (the process evaluation plans are part of the overall EDC evaluation plan).
[67] New York Department of Public Service. 2016. Evaluation, Measurement & Verification Guidance. Weblink

and reported in real time. When evaluators work with program designers and managers during program development and embed the evaluation into the program, data can be collected throughout the implementation period that informs the program staff about opportunities for improvement. Real-time evaluations typically last for one to two years, with ongoing data collection and quarterly to bi-annual reporting that targets the type of information program staff needs to gauge their program's progress and effectiveness.

3.5.2 Data Collection and Evaluation Activities

Process evaluation efforts can include a wide range of data collection and assessment efforts, including:

- Interviews and surveys with an EDC's program designers, managers, and implementation staff (including contractors, sub-contractors, and field staff)
- Interviews and surveys with trade allies, contractors, suppliers, manufacturers, and other market actors and stakeholders
- Interviews and surveys with participants and non-participants
- Interviews and surveys with people using the technologies (e.g., usability studies of websites)
- Interviews and surveys with key policymakers
- Observations of operations and field efforts, including field tests and investigative efforts
- Operational observations and field-testing, including process-related M&V efforts
- Workflow, production, and productivity measurements
- Reviews, assessments, and testing of records, databases, program-related materials, and tools
- Collection and analysis of relevant data or databases from third-party sources (e.g., equipment vendors, trade allies and stakeholders, and market data suppliers)
- Focus groups with participants, non-participants, trade allies, and other key market actors associated with the program or the market in which the program operates.

Data collection for process evaluations may also include acquisition of information that is used for impact evaluations (e.g., free ridership and spillover information to help estimate net savings). The following sections describe in more detail considerations to be followed in data collection.

3.5.2.1 Review of Program Information and Data

Process evaluators glean a wealth of information about the program from information and records that the program maintains, including the tracking system; program communications documents (usually electronic); and the materials used for marketing, outreach, and publicity. There may also be process flow diagrams, program theory and logic documents, planning documents, and regulatory documents that set forth the purpose and intention of the program. The process evaluator should be familiar with these documents, using them to understand the context for the program and to provide data in addition to those obtained in interviews.

3.5.2.2 Interviews with Program Managers, Administrators, and Implementers

Program managers and staff are an essential source of information, as they typically know the program better than anyone. Interviews with lead program planners and managers, their supervisors, and a sampling of program staff, including both central staff and field staff, is the first step in a process evaluation. Data from these interviews help the evaluator assess the program design and operations to recommend any changes to improve the program's ability to obtain cost-effective energy savings.

Subjects important to discuss with these individuals include overall understanding of program goals and objectives, available and needed resources for program implementation, program impact on the market, communication within the program, communication with customers and stakeholders, and barriers to program administration and participation. In addition, through the interviews, evaluators can get a sense of the program's strengths and weaknesses, its successes, and the quality of work; they can then compare this information with information stakeholders and participants express during interviews and surveys.

3.5.2.3 Interviews, Surveys, and Focus Groups with Key Stakeholders and Market Actors

In addition to program staff, many other individuals are involved in a program, including policymakers (such as PUC staff); utility managers; key stakeholders (including trade associations and tenant groups); and other market actors, such as product manufacturers, distributors, installation contractors, and service personnel. It is useful to interview a sample from a variety of key market actor groups to obtain their insights into the program's impact on the market, what it is doing well, and what can be improved.

3.5.2.4 Interviews, Surveys, and Focus Groups with Participants and Non-participants

One purpose of virtually all process evaluations is to understand the customer's experience to inform program improvements. Program participants have valuable perspectives on aspects of the program that work well and others that represent barriers to participation or satisfaction. Detailed feedback from participants is also important for determining whether the customer's perceptions of specific program attributes and delivery procedures conflict or mesh with those of program designers and managers. Beneficial detailed feedback can include levels of satisfaction with various elements of the program, such as the product(s), organization, scheduling, educational services, quality of work performed, attitude of site staff, responsiveness to questions/concerns, and savings level achieved.

3.5.2.5 Other Types of Data Collection Efforts

There are many other types of data collection methods to consider, including ride-along observations with auditors or contractors; intercept surveys; mystery shopping; shelf-stocking counts; and electronic, in-person, or mail data collection instead of phone surveys. Similar data to those mentioned above, if collected for programs in other jurisdictions, can be used to draw comparisons or develop best practices. It is essential to select the optimal data collection approach and the appropriate sample, and to draw conclusions consistent with the limits of the data and sample.

3.5.3 Process Evaluation Analysis Activities

The process or market evaluation analysis is considered triangulation. Because much of the data is qualitative, the evaluation team's analysts must be systematic and careful to draw accurate conclusions across the different sources.

Evaluators must construct the data collection instruments carefully to ensure that similar questions are posed across groups; it is also essential to select samples that accurately represent the target audience so that the evaluator's conclusions are justified.

3.5.4 Process and Market Evaluation Reports

Each process evaluation should include the findings from the research tasks and provide conclusions and recommendations that address the research objectives. The EDC, SWE, and the PUC cannot implement long lists of recommendations. Instead, a short list of targeted, actionable recommendations and the status of the recommendations is expected.

3.6 Sampling Statistics and Presentation of Uncertainty

Gross verified energy and demand savings estimates for EE&C programs are usually determined through the observation of key measure parameters among a sample of program participants. A census evaluation would involve surveying, measuring, or otherwise evaluating the entirety of projects within a population. Although a census approach would eliminate sampling uncertainty, the reality is that EM&V takes many resources, so sampling is needed. When a representative sample of measures, projects, or participants is selected and analyzed, the sample statistics provide a reasonable estimate of the population parameters.

There is an inherent risk associated with sampling because, even with the best sample design, the projects selected in the evaluation sample may not be representative of the program population with respect to the parameters of interest. Sample sizes affect the uncertainty of the resulting estimates. Typically, as the proportion of projects in the program population that are sampled increases, the sampling uncertainty decreases because we have information about a greater number of population units. The amount of variability in the population and sample also affects the uncertainty. A small sample drawn from a homogeneous population will provide a more reliable estimate of the true population characteristics than a small sample drawn from a heterogeneous population. Variability is expressed using the coefficient of variation (Cv) for programs that use simple random sampling and an error ratio for programs that use ratio estimation. The Cv of a population is equal to the standard deviation (σ) divided by the mean (μ), as shown in Equation 2.

Equation 2: Coefficient of Variation

$$\boldsymbol{C_v} = \frac{\boldsymbol{\sigma}}{\mu}$$

When ratio estimation is utilized, the ratio of verified savings to reported savings can vary for each unit in the sample. For sampling and precision purposes, we are interested in how the unit-level ratios compare to the overall ratio for the sample. Are they consistent or highly variable? The error ratio is an expression of this variability and is analogous to the Cv for simple random sampling.

Equation 3 provides the formula for estimating error ratio.[68] The sampling unit will vary depending on program design, how participation is tracked, and the segmentation approach used by the EC. In this section, we use projects as the sampling unit, but in practice the sampling unit may be distinct participants, rebate applications, groupings of like measures installed by a participant, or some other definition. EDC ECs should clearly define the sampling unit in their Evaluation Plans and sample design memos. The Ω term in Equation 3 is equal to the difference between the project-level verified savings estimate (γ) and the realization rate multiplied by the reported savings.

Equation 3: Error Ratio

$$\boldsymbol{Error\ Ratio} = \frac{\sum_{i=1}^{N} \boldsymbol{\Omega_i}}{\sum_{i=1}^{N} \boldsymbol{\gamma_i}}$$

Equation 4 shows the formula used to calculate the required sample size for an evaluation sample[69] based on the desired level of confidence and precision. Notice that the Cv term is in the numerator, so required sample size will increase as the level of variability increases.

Equation 4: Required Sample Size

$$n_0 = \left(\frac{z * C_v}{D}\right)^2$$

Where:

n_0 = The required sample size before adjusting for the size of the population

Z = A constant based on the desired level of confidence (equal to 1.645 for 90% confidence, two-tailed test)

C_v = Coefficient of variation (standard deviation/mean)

D = Desired relative precision

Unfortunately, the EC does not know the C_v or error ratio values until after the verified savings analysis is complete and thus must make assumptions about the level of variability in the savings values based on previous program years or evaluations of similar programs in other jurisdictions. In the absence of prior information regarding the C_v for the targeted population, EDC ECs can assume a default C_v equal

---

[68] Equation 3 is based on the methodology set forth in the California Evaluation Framework. The National Renewable Energy Laboratory's (NREL) UMP provides a slightly different formula for the calculation of error ratio that is an acceptable alternative if ECs wish to use it.

[69] If ratio estimation is used, evaluators may replace C_v with error ratio in Equation 4.

to 0.5 for each sample population to determine target sample sizes. Regarding the default assumption of C_v of 0.5, the Sample Design UMP[70] states:

> *One reason C_v of 0.5 are often reasonable in evaluation work is that the savings values are typically positive for all (or nearly all) projects. If 95% of a program's projects have savings between zero and 200% of the mean savings, and if the savings values are approximately normally distributed, then a CV of 0.5 will apply.*

Once the CV has been measured, evaluators may use that historical CV in developing their sampling plans. Evaluators should estimate the CV values for each sampled population and report the values in their final annual reports so they can be used in subsequent evaluation plans.

The sample size formula shown in Equation 4 assumes that the population of the program is infinite or large. In practice, this assumption is not always met.

> For sampling purposes, any population greater than approximately 7,000 may be considered infinite. No adjustment is required in this case, and the final sample size can be calculated using Equation 3. For smaller, finite populations, the use of a finite population correction factor (FPC) is warranted. This adjustment accounts for the decreases in uncertainty that result when the number of sampled projects is a large proportion of the smaller population. Multiplying the results of Equation 4 by the FPC formula shown in Equation 5 will produce the required sample size for a finite population.

Equation 5: Finite Population Correction Factor

$$fpc = \sqrt{\frac{N - n_0}{N - 1}}$$

Where:

N = Size of the population

n_0 = The required sample size before adjusting for the size of the population

The required sample size (n) after adjusting for the size of the population is given by Equation 6.

Equation 6: Application of the Finite Population Correction Factor

$$n = n_0 * fpc$$

[70] Uniform Methods Protocols: Chapter 11: Sample Design Cross-Cutting Protocols at page 53. Weblink

3.6.1 Evaluation Precision Requirements

Table 16Table provides minimum levels of sampling uncertainty prescribed for the Act 129 gross impact evaluations to balance the need for accurate savings estimates while limiting the costs of evaluation. The values in Table 16 apply to both energy and peak demand and assume a two-tailed design and specify the relative precision that must be met or exceeded at the given confidence level each time a gross impact evaluation is conducted. The values in Table 16 are also suggested for NTG and process evaluations but are not a requirement like they are for gross impact evaluations. See Section 4.5.2 for more details pertaining to process evaluation sampling. Note there are no minimum precision requirements for EDC evaluations of Phase V savings at the portfolio level. However, if the minimums established in Table 16 are met for independent impact evaluations, the relative precision values of the total Phase V savings will meet or exceed the requirements at the same levels of confidence.

Table 16: Minimum Confidence and Precision Levels

Portfolio Segment	Confidence and Precision Level
Residential Portfolio	90/10
Non-residential Portfolio	90/10
Individual Initiatives Within Each Portfolio	85/15

The definition of the term initiatives in Table 16 is important and has clear implications for sample design and allocation of resources. Delivery channel is the most important characteristic, but EDCs and their ECs may also wish to consider the targeted end-use or other characteristics when defining initiatives for evaluation purposes. In some cases, an initiative will be the same as a program in an EDC's EE&C plan. In other words, some programs are composed of a single initiative, and the initiative is only offered in a single program. However, other Phase V programs, as defined in approved EE&C plans, include multiple initiatives that should be evaluated separately. For example, an EE&C plan may include a large residential energy efficiency program composed of rebates for efficient equipment, kits of measures distributed via mail, and appliance recycling. These are three distinct initiatives that should be sampled and evaluated separately with each initiative subject to the precision requirements in Table 16. Initiatives may also span multiple programs. For example, an EE&C plan may include a small C&I program and a large C&I program that both include prescriptive lighting rebates. ECs may elect to define prescriptive lighting as an initiative and combine projects from multiple programs into a single evaluation sample if the project population is expected to be homogeneous and historical realization rates have been steady for the initiative.

The SWE recommends that ECs submit a memo to the SWE for approval that outlines the definition of evaluation initiatives and the impact evaluation cadence prior to drafting a complete EM&V plan. Section 3.8 provides additional detail regarding the frequency of impact evaluations.

Special consideration should be given to the following situations:

1. Crosscutting initiatives that span both the residential and non-residential sectors must[71] be evaluated separately (i.e., one evaluation for the residential sector and one evaluation for the non-residential sector).
2. ECs may choose to define evaluation initiatives in a way that includes both residential low-income and residential non-low-income projects. In this scenario, the two sectors should be treated as distinct strata with results calculated and reported separately, but precision requirements from Table 16 do not need to be achieved for each sector. The 85/15 requirement applies to the initiative as a whole.
3. The non-residential sector evaluation should include no fewer than three initiatives. The list below provides suggestions for possible definitions of initiatives within the non-residential portfolio.
    a. Prescriptive Lighting
    b. Prescriptive Non-Lighting
    c. Custom rebates
    d. Direct installation
4. The residential sector evaluation should include no fewer than four initiatives. Within the residential portfolio, a potential group of initiatives might be:
    a. Audits and weatherization / Whole-house program
    b. Appliance Recycling
    c. School education and other kit offerings
    d. Rebates for efficient products
5. It often is more challenging to obtain accurate peak demand savings estimates than annual energy savings estimates, and peak demand savings estimates can exhibit a greater degree of variability between ex ante and ex post. The minimum levels of precision established in Table 16 are required for both energy and peak demand savings estimates. EDC ECs should consider the expected Cv for energy and demand separately and design samples around the parameter with higher expected variability.

ECs may use their professional judgment in the design of the sample as long as they meet the minimum precision requirements. ECs should design evaluation samples to exceed the minimum requirements so they will not miss the precision requirements established in this Evaluation Framework if program characteristics (population size, variability) are slightly greater than anticipated. If the confidence and precision targets are not met, corrective actions will be required in the current or subsequent impact evaluation year within the compliance period. For Phase V, EDC ECs are encouraged to rotate impact evaluation activities so that not every initiative receives an impact evaluation all five years of the phase. In certain cases, EDCs and their ECs will be permitted to use historic realization rates to determine

71 The SWE may approve exceptions during the review of EDC EM&V plans. For example, small businesses may be eligible to participate in an appliance recycling program, but 99% of the program savings will come from the residential sector. The 1% of program savings from the non-residential sector does not need to be evaluated as a standalone program.

verified savings for program years when they do not complete an impact evaluation. However, impact evaluations that fail to meet the minimum precision requirements are not permitted to be used as historic realizations rates.

It is important to note that the requirements in Table 16 are for relative precision. When realization rates are low, gross verified savings fall short of projections and the relative precision of the results is likely to be poor. If precision targets are missed primarily because of a low realization rate, the SWE will take this into account during audit activities and findings will focus on correcting the underlying issue as opposed to modification of the sample design.

ECs are encouraged to use stratification to ensure that the sample is efficiently designed. Evaluators should use their professional judgment to develop size thresholds and definitions for the project strata, subject to review and approval by the SWE. The SWE audit of evaluator sample designs is discussed in more detail in Section 3.9. For high-impact or high-uncertainty project strata, evaluators should ensure that they evaluate savings using an enhanced level of rigor.

Programs such as low-income weatherization, behavior modification, retro commissioning, strategic energy management, or customer education often rely on a billing regression analysis of a census or near census of program participants to determine verified savings. These programs require special consideration because a census, rather than a sample, of program participants is evaluated, so theoretically there is no sampling uncertainty. Instead, the precision of savings estimates is determined using the standard error of the regression coefficient(s) that determine savings. Depending on program size and the magnitude of per-participant savings, the requirements in Table 16 may not be feasible for programs that use a census regression approach.

The SWE has established specific requirements for behavioral programs in Section 6.1. For other programs that use a billing regression analysis, the precision requirement is essentially statistical significance. If the 85% confidence interval around the savings estimates includes 0 kWh, an EDC should explain remedial actions that will be taken to improve the precision of the savings estimate. For example, if the per-home savings estimate for a program is equal to 200 kWh/year ± 400 kWh/year, remedial actions should be taken in the same program year or the following program year to improve the precision of the savings estimate. If it is not possible to achieve more precise results using billing regression analysis, the EDC EC should explore an alternative measurement technique for future impact evaluations.

3.6.2 Overview of Estimation Techniques

Evaluators may choose to employ two broad classes of probability estimation techniques in the impact evaluation of EE&C programs.

1. **Mean-per-unit estimation** (estimation in the absence of auxiliary information): This technique is useful if the projects within a population are similar in size and scope. Simple random sampling is recommended for residential programs that include many rebates for similar equipment types.

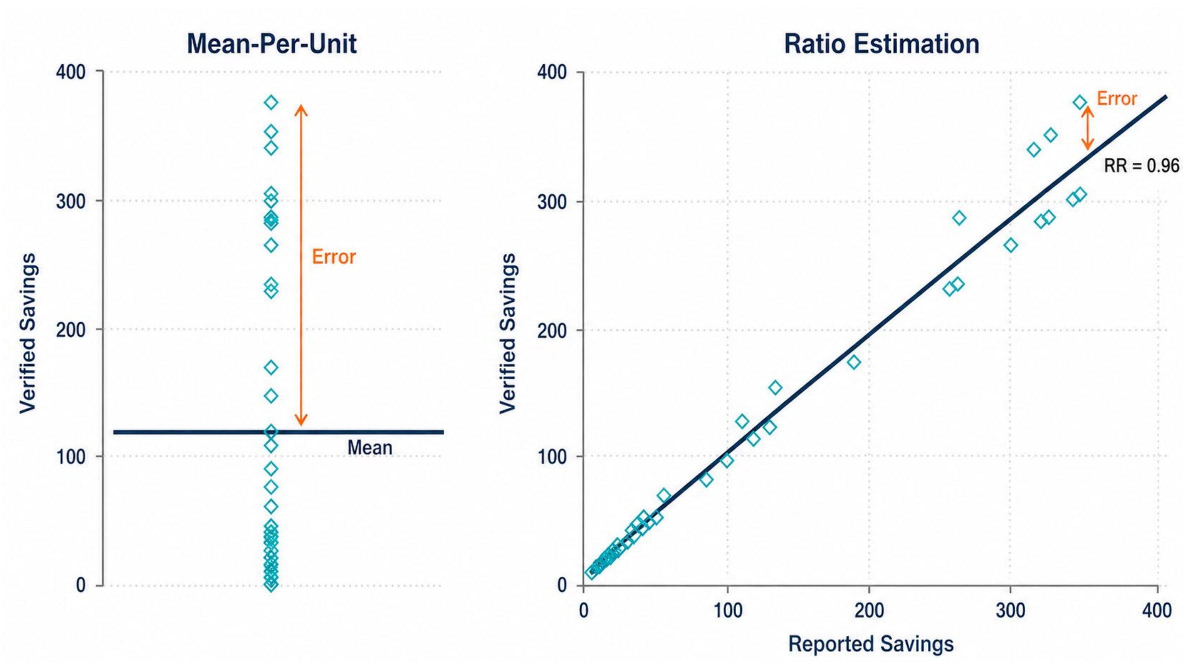
2. **Ratio estimation** (estimation using auxiliary information): This is recommended for non-residential programs, or residential programs offering a variety of measures with varying savings, because the sizes of the savings estimates of the projects within a program vary considerably within the program population. Ratio estimation can be used with or without stratification. This technique relies on auxiliary information reported in the program tracking system – usually the ex-ante kWh and kW savings of the projects. This technique assumes that the ratio of the sum of the verified savings estimates to the sum of the reported savings estimates within the sample is representative of that ratio for the program population. This ratio is referred to as the realization rate, or ratio estimator, and is calculated as follows:

$$Realization\ Rate = \frac{\sum_i^n Verified\ Savings}{\sum_i^n Reported\ Savings}$$

where n is the number of projects in the evaluation sample.

Figure 6 shows the reduction in error that can be achieved through ratio estimation when the sizes of projects within a program population vary considerably. The ratio estimator can provide a better estimate of individual project savings than a mean savings value by leveraging the reported savings estimate.

Figure 6: Comparison of Mean-Per-Unit and Ratio Estimation



Sample stratification can be used with either of the two classes of estimation techniques presented previously. Stratified random sampling refers to the designation of two or more sub-groups (strata) from within the program population prior to the selection process. It is imperative that each sampling unit (customer/project/measure) within the population belongs to one (and only one) stratum. Typically, the probability of selection is different between strata; this is a fundamental difference from simple random sampling, where each sampling unit has an identical likelihood of being selected in the sample. The inverse of the selection probability is referred to as the case weight and is used in estimation of impacts when stratified random samples are utilized. Stratification is advantageous for the following reasons:

- Increased precision if the within-stratum variability is small compared to the variability of the population as a whole. Stratification potentially allows for smaller total sample sizes, which can lower evaluation costs.
- A stratified sample design allows ECs to ensure that a minimum number of units within a particular stratum will be verified. For example, a C&I program with 1,000 projects in the population, may only have ten that are CHP projects. If the sample size is 40 and simple random sampling is used, each project has a 4% chance of being included in the sample, and the probability that the resulting sample contains one or more CHP projects is only 33.6%. On the other hand, if stratified random sampling is used and one stratum is defined as including only CHP projects, then as long as the sample size within each stratum is one or more projects, the sample will include a CHP project with certainty and each CHP project will have a 10% probability of being selected.
- Additional sample designs can be considered within each stratum. It is easy to implement a value-of-information approach through which the largest projects are sampled at a much higher rate than smaller projects.
- Sampling independently within each stratum allows for comparisons among groups. Although this Framework only requires that a single relative precision be met at the program level annually, EDCs and their ECs may find value in comparing results between strata (e.g., comparing the verification rates between measures within a program).

ECs are encouraged to limit the use of simple random sampling to evaluations with homogenous measure populations, such as Appliance Recycling, and to employ stratification for initiatives which offer a diverse mix of measures. However, the choice of using stratified random sampling or simple random sampling is ultimately left up to the discretion of the EDC EC.

3.6.3 Presentation of Uncertainty

In the final annual report for each program year, each EDC should report the verified energy and demand savings achieved by each program in its portfolio and estimates for the entire portfolio. Verified savings estimates should always represent the point estimate of total savings, or the midpoint of the confidence interval around the verified savings estimate for the program. In addition to the verified savings estimates for energy and demand, EDCs should report the error bound, or margin of error, and the relative precision of the savings estimate such that:

Equation 7: Error Bound of the Parameter Estimate

$$Error\ Bound = se * (z\text{-}statistic)$$

where:

se = The standard error of the estimated population parameter of interest (proportion of customers installing a measure, realization rate, total energy savings, etc.). This formula will differ according to the sampling and estimation techniques utilized.

$z\text{-}statistic$ = Calculated based on the desired confidence level and the standard normal distribution.

Table 17 provides the appropriate z-statistic to use for several commonly used confidence levels. Each value assumes a two-tailed design.

Table 17: Z-statistics Associated with Common Confidence Levels

Confidence Level	Z-statistic
80%	1.282
85%	1.440
90%	1.645
95%	1.960

Use of a z-statistic implies normality. The Central Limit Theorem shows that the means of sufficiently large random samples drawn from a population will follow a normal distribution, even if the population that is the source of the sample is not normally distributed. However, for sample sizes smaller than 30, the Central Limit Theorem begins to break down and the normality assumption is no longer valid. A t-distribution is the appropriate distribution for evaluators to consider when drawing samples of fewer than 30 projects/measures. In this case, a t-statistic will be used in estimation once the sample has been collected. The t-statistic replaces the z-statistic in Equation 7 and is calculated using the degrees of freedom (sample size minus the number of estimates). As the sample size becomes larger, the t-statistic gets closer to the z-statistic.

In cases where the parameter of interest is a proportion or realization rate, the estimate is applied to the reported savings values to calculate the gross verified savings for the program. The error bound of the verified savings estimate (in kWh/year or kW) should be reported for each program and is calculated as follows:

Equation 8: Error Bound of the Savings Estimate

$$Error\ Bound_{(kWh\ or\ kW)} = Error\ Bound_{Parameter} * Gross\ Reported_{(kWh\ or\ kW)}$$

The relative precision value of the verified savings estimate for each program should be reported, as well as the confidence level at which it was calculated. This formula is shown in Equation 9:

Equation 9: Relative Precision of the Savings Estimate

$$Relative\ Precision_{Verified\ Savings} = \frac{Error\ Bound_{(kWh\ or\ kW)}}{Gross\ Verified_{(kWh\ or\ kW)}}$$

Evaluations of programs that use stratified ratio estimation require an additional step because each stratum will have its own realization rate and error bound that should be reported.

At the conclusion of Phase V of Act 129, each EDC will have five verified savings estimates for energy and five verified savings estimates for demand for each initiative in its portfolio. The Phase V verified savings estimate is the sum of these values. These verified savings estimates will be calculated through as many as five impact evaluations. Although the error bound estimates for each impact evaluation are expressed in the unit of interest (kWh/year or kW), they cannot be summed to produce the error bound for Phase V impacts. Equation 10 shows the formula for calculating the error bound of the Phase V impacts for a program that receives two impact evaluations: one for PY18 and PY19 and a second for PY20-PY22. The same methodology should be used to calculate the error bound and relative precision of the annual sector- and portfolio-level verified savings estimates. Phase V error bounds and relative precisions should be calculated and reported at the 90% confidence level. This will require a recalculation of the annual error bounds if the 85% confidence level were used for a program. To convert the annual error bound to the 90% confidence interval, evaluators should perform the calculations shown in Equation 7 and Equation 8 using the standard error of the parameter estimate and the z-statistic associated with the 90% confidence interval (1.645).

Equation 10: Phase V Error Bound

$$Error\ Bound_{Phase\ V} = \sqrt{Error\ Bound_{PY18,PY19}^2 + Error\ Bound_{PY20-PY22}^2}$$

Using this methodology, evaluators will have a Phase V verified savings estimate for the initiative and an error bound for that estimate. The relative precision of the Phase V verified savings for the program is then calculated using these two values.

Equation 11: Relative Precision of Phase V Savings Estimate

$$Relative\ Precision_{Phase\ V} = \frac{Error\ Bound_{Phase\ V}}{Gross\ Verified\ Savings\ Estimate_{Phase\ V}}$$

Equation 10 also should be used to combine the Phase V error bounds from programs to the sector level and from the sector level to the portfolio level. Note that Equation 10 assumes that estimated savings in each impact evaluation are independent. The independence assumption must hold for this formula to be applied to the combination of program-level savings to the sector level within a portfolio and/or program year.

3.6.4 Systematic Uncertainty

Section 3.6.1 of the Evaluation Framework discusses the uncertainty that is introduced into evaluation findings when a sample, rather than a census, of projects is used to determine program impacts. Sampling uncertainty, or error, is largely random and can be estimated using established statistical procedures. On the other hand, systematic uncertainty represents the amount of error that is introduced into evaluation results consistently (not randomly) through the manner in which parameters are measured, collected, or described. Systematic uncertainty is more challenging to quantify and mitigate than sampling uncertainty because sources of systematic uncertainty often are specific to the program, measure, or site being evaluated. However, to present evaluation results as though sampling error is the only source of uncertainty in an evaluation misrepresents the accuracy with which an EDC can estimate the impacts achieved by its EE&C Plan. EDC final annual reports should discuss major sources of systematic uncertainty and the efforts the EC made to mitigate them.

Common sources of systematic uncertainty, which should be considered in an EDC's evaluation plan include:

1. **Deemed or Stipulated Values** – TRM values are based on vetted engineering principles and provide reasonable estimates of measure energy and demand impacts while expending relatively few evaluation resources. Using these values in evaluation results can introduce considerable bias if the values are not adequately prescribed or do not fully capture the complexity of a measure. Dated values or adjusted values from secondary research are likely to introduce systematic error in the evaluation findings.
2. **Data Collection and Measurement** – According to sampling theory, when a project is selected in the impact evaluation sample and energy and demand savings values are calculated, those savings values are discrete. In reality, the reliability of these estimates is subject to a host of uncertainties that must be considered. Survey design can introduce a variety of biases into evaluation findings. Consider a lighting survey that includes questions to a facility contact about the typical hours of operation in their building. If the survey does not include questions about business closings for holidays, the survey responses will systematically overestimate the HOU of fixtures in the facility. Evaluators also must consider another source of systematic uncertainty, human error. If the engineer visiting a site in the evaluation sample forgets to complete a key field on the data collection instrument, an assumption must be made by the analyst calculating savings for the project regarding the parameter in question. On-site metering is considered a high-rigor evaluation approach and is reserved for high-impact/high-uncertainty projects, but these results can be biased by equipment placement, poor calibration, or differences in the pre/post metering period not addressed in the analysis.
3. **Sample Design** – Evaluation samples are constrained by evaluation budgets and the practicality of collecting information. Non-coverage errors can arise if the sample does not accurately represent the population of interest. For instance, an evaluation survey that is conducted via email with a random sample of EDC customers necessarily excludes all customers who do not have an email address or have chosen not to provide their EDC with this information. If this population of customers somehow differs from the population of customers with known email addresses (the sample pool) with respect to the parameter in question, the value calculated from the sample will not accurately reflect the population of interest as a whole.
4. **Participation Likelihood** – Non-response and self-selection errors occur when some portion of the population is less likely (non-response) or more likely (self-selection) to participate in the evaluation than other portions. Retired customers frequently are over-represented in residential evaluation findings because daytime recruiting calls to a home phone number are far more likely to reach retired program participants. Values calculated from samples that over-represent certain segments and under-represent others are subject to systematic uncertainty if the customer segments differ with respect to the parameter of interest.

The systematic uncertainty resulting from data collection and measurement, or sample design cannot be easily quantified with a formula. EDC evaluators should discuss the steps taken to mitigate systematic error from these sources and any analysis undertaken to understand where significant sources may exist. The UMP Sample Design Protocol[72] identifies six areas which may be examined to determine how rigorously and effectively an evaluator has attempted to mitigate sources of systematic error. A summary of the six areas is as follows:

1. Were measurement procedures (such as the use of observational forms or surveys) pretested to determine if sources of measurement error could be corrected before the full-scale fielding?
2. Were validation measures (such as repeated measurements, inter-rater reliability, or additional subsample metering) used to validate measurements?
3. Was the sample frame carefully evaluated to determine which portions of the population, if any, were excluded in the sample? If so, what steps were taken to estimate the impact of excluding this portion of the population from the final results?
4. Were steps taken to minimize the effect of non-response or self-selection in surveys or other data collection efforts? If non-response appears to be an issue, what steps were taken to evaluate the magnitude and direction of potential non-response bias? Were study results adjusted to account for non-response bias via weighting or other techniques?[73]
5. Has the selection of formulas, models, and adjustments been conceptually justified? Has the evaluator tested the sensitivity of estimates to key assumptions required by the models?
6. Did trained, experienced professionals conduct the work? Was the work checked and verified by a professional other than the one conducting the initial work?

EDC evaluation plans and final annual reports should discuss the steps ECs took to answer as many of the questions above as possible in the affirmative. SWE audit activities will consider the appropriateness of evaluators' techniques to mitigate systematic uncertainty and identify areas where changes or additional research are warranted.

3.6.5 Additional Resources

The 2009 and 2011 versions of the Audit Plan and Evaluation Framework for Pennsylvania Energy Efficiency and Conservation Programs include detailed information regarding sample design, sample size calculations, definitions and formulas for error ratio, CV, and relative precision. This information has been excluded from subsequent versions of the Evaluation Framework. If EDCs, their ECs, or stakeholders require additional information regarding sampling, the following resources will be helpful:

---

[72] The protocols can be found here.

[73] Some common methods to deal with non-response by incorporating response rates into the sampling weights are presented in *Applied Survey Data Analysis* by Heeringa, West, and Berglund (2010).

- The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. Prepared for the National Renewable Energy Laboratory by The Cadmus Group, January 2013.
- The California Evaluation Framework. Prepared for the California Public Utilities Commission and Project Advisory Group by TecMarket Works, June 2004.
- Audit Plan and Evaluation Framework for Pennsylvania Act 129 Energy Efficiency and Conservation Programs. Prepared for the PUC by GDS Associates, November 2011.

3.7 Cost-Effectiveness

Verified gross and verified net results from the EDCs' evaluation activities will be input into a benefit-cost model to assess the cost-effectiveness of the EDCs' efforts at the initiative, sector, and portfolio levels. In accordance with the PUC's requirements for determining cost-effectiveness, the EDC's EE&C programs will be evaluated based on the TRC Test. The guidelines for the Phase V TRC are stipulated in the 2026 TRC Test Final Order. All cost-effectiveness evaluations and assessments will be conducted in accordance with the PUC's latest TRC Test Order.

3.7.1 TRC Method

The 2026 TRC Test Final Order builds on all previous Act 129 TRC Test orders and industry documents, such as the CaSPM,[74] for the benefit-cost analysis of EE&C plans for Phase V. Act 129 defines the TRC Test as "a standard test that is met if, over the effective life of each plan not to exceed 15 years, the net present value of the avoided monetary cost of supplying electricity is greater than the net present value of the monetary cost of energy efficiency conservation measures."[75]

Since its update in 2002,[76] the CaSPM manual has served as the basis for cost-effectiveness testing in virtually every state with energy efficiency programs, including Pennsylvania. According to the CaSPM:

> *The Total Resource Cost Test measures the net costs of a demand-side management program as a resource option based on the total costs of the program, including both the participants' and the utility's costs. The test is applicable to conservation, load management, and fuel substitution programs. For fuel substitution programs, the test measures the net effect of the impacts from the fuel not chosen versus the impacts from the fuel that is chosen as a result of the program. TRC Test results for fuel substitution programs should be viewed as a measure of the economic efficiency implications of the total energy supply system (gas and electric).*
>
> *Benefits and Costs: This test represents the combination of the effects of a program on both the customers participating and those not participating in a program.*

[74] The California Standard Practice Manual – Economic Analysis of Demand-Side Programs and Projects, July 2002, Weblink
[75] Act 129 of 2008 – House Bill 2200. Weblink Page 61
[76] Economic Analysis of Demand-Side Programs and Projects. Weblink, Page 18.

The CaSPM framework was the starting point for the Pennsylvania TRC Test but it does not address all issues specific to Pennsylvania. The TRC Test for Phase V of Act 129 differs in several notable ways from the TRC Test as described in the CaSPM. The PUC, in collaboration with stakeholders through formal public comment proceedings, has customized the TRC Test to reflect Pennsylvania-specific policies and priorities. For Phase V, the Commission designed the 2026 TRC Test Final Order (issued November 7, 2024) to provide all instructions for Act 129 cost-effectiveness testing in a single, comprehensive document, leveraging insights from multiple resources, including the CaSPM and previous Act 129 TRC Test Orders. EDC ECs should refer to the 2026 TRC Test Final Order for Phase V for detailed formulae and definitions related to the proper calculation of the Act 129 TRC Test.[77]

3.7.2 Application of Avoided Costs

For Phase V, the Commission proposed continued use of the status quo Act 129 methodology to forecast the avoided costs of supplying electricity. The Commission believes that more detailed instructions improve consistency across EDCs and leads to better alignment with market conditions. To that end, the Phase V SWE developed an updated MS-Excel spreadsheet calculation model (ACC[78]) to implement the methodology outlined in the Final Order. EDCs must utilize this standard tool when developing avoided costs for Phase V. The calculation methodology for each cost category is described in Table 18 below briefly and more detailed descriptions can be found in the 2026 TRC Test Final Order.

Table 18: Phase V Avoided Cost Calculation Methodology

Avoided Cost Category	Phase V Methodology
Electric Energy	Continue to use a 20-year period, dissected into three segments. Costs will continue to be calculated in a time-differentiated format, with six distinct periods per annum. The cost sources for each segment are described below: Segment 1 (2026–2029: NYMEX electricity futures prices at the PJM Interconnection Western Hub location with an EDC zonal basis adjustment Segment 2 (2030–2035): Medium-term NYMEX natural gas futures blended with U.S. Energy Information Administration's (EIA's) Annual Energy Outlook project natural gas costs, converted to electric energy price with a spark price spread calculation Segment 3 (2036–2045): Long-term EIA Annual Energy Outlook projected natural gas costs, converted to electric energy price using a spark price spread calculation
Generation Capacity	Use actual zonal PJM BRA clearing prices when available. When actual prices are not available, future costs can be projected using the five most recent BRA clearing prices for the zone escalated using a compound annual growth rate of the BLS index for the power sector and then apply the inflation rate (2%).

[77] Pennsylvania Public Utility Commission, *2026 Total Resource Cost Test Order*, Docket No. M-2024-3048998, November 7, 2024.
[78] Weblink

Avoided Cost Category	Phase V Methodology
Transmission & Distribution Capacity	The Phase IV SWE, at the direction of commission, conducted a T&D study in collaboration with EDC planners to develop a granular forecast of avoided T&D costs by EDC in preparation for Phase V. The study employed probabilistic deferral methodologies combined with actual T&D costs from each EDC to estimate the avoided T&D cost values. The ACC tool provides EDCs with avoided T&D values for each year of Phase V. For program participants who take service at primary voltage, only the avoided cost of transmission capacity is applied.
Compliance with Alternative Energy Portfolio Standards Act (AEPS)	Using Marex[79] forecast data from May 1, 2024, for the year 2026, the Commission directed that the AEPS Act avoided costs be $6.88 per MWh for the first year of Phase V and escalated by the BLS escalation factor and the 2% inflation rate every year thereafter.
Price Suppression Effects	In preparation for the 2026 TRC Test Tentative Order and Phase V planning, the SWE completed an analysis of price suppression effects using data on the locational marginal price of energy and PJM BRA auction results and developed a recommended set of price suppression effect values. The price suppression effects to be used for Phase V only include the benefits that accrue to Pennsylvania ratepayers.
Reduction in Arrearages and Collection Costs	The SWE conducted a study to quantify and monetize EDCs' financial savings through an analysis of EDC data on customer arrearages, shutoffs, and collection actions for Act 129 low-income program participants. The estimates avoided cost values associated with this benefit stream are EDC specific, considering differences in the total costs associated with arrearages and collections at each EDC.

For some EE&C measures, a single baseline may not be appropriate for the duration of the mechanical life of the equipment. Although compliance is based on first-year savings, lifetime savings are required for the calculation of TRC benefits. Dual baselines may be appropriate for early replacement measures when the existing equipment that serves as the baseline initially is expected to reach the end of its useful life before the efficient measure and a code-minimum baseline needs to be assumed for the remainder of the measure life. EDCs and their ECs are expected to utilize dual baselines where appropriate and practical.

3.7.3 Aligning Measure Savings with Incremental Measure Costs

To determine energy efficiency cost-effectiveness using the TRC Test, the EE&C measure/program savings and costs must be determined and aligned properly. For the TRC Test, the appropriate cost to use is the cost of the energy efficiency device in excess of what the customer otherwise would have spent, regardless of what portion of that incremental cost is paid by the participant or paid by an EDC. Thus, the incremental measure cost (IMC) should be evaluated with respect to a baseline. For instance, a program that provides an incentive to a customer to upgrade to a high-efficiency central air conditioner would use the cost difference between the efficient air conditioner and the code minimum baseline model. Similarly, the savings are calculated as the reduced energy consumption of the efficient unit compared to the baseline model.

---

[79] Marex is a United Kingdom-based broker of financial instruments and provider of market data services across the metals, agricultural and energy markets. Weblink

Table 19 lists five basic measure decision types, along with a summary of the definition of IMCs and savings for each of the decision types.

Table 19: Measure Decision Types

Type of Measure	IMC ($/Unit)	Impact Measurement (kWh/year/Unit)
New Construction	Cost of efficient device minus cost of baseline device	Consumption of baseline device minus consumption of efficient device
Replace on Burnout (ROB)	Cost of efficient device minus cost of baseline device	Consumption of baseline device minus consumption of efficient device
Retrofit: An additional piece of equipment or process is retrofit to an existing system. (e.g., additional insulation or duct sealing)	Cost of efficient device plus installation costs	Consumption of old device minus consumption of efficient device
Early Replacement: Replacement of existing functional equipment with new efficient equipment	Present value of efficient device (plus installation costs). If a dual baseline is used, subtract the present value of baseline device assumed to be installed in at the end of remaining useful life of the existing equipment (plus installation costs)	During remaining life of old device: Consumption of old device minus consumption of efficient device After remaining life of old device: Consumption of baseline device minus consumption of efficient device
Early Retirement (No Replacement)	Cost of removing old device	Consumption of old device

* The early replacement case is essentially a combination of the simple retrofit treatment (for the time period during which the existing measure would have otherwise remained in service) and the failure replacement treatment for the years after the existing device would have been replaced.

The 2026 TRC Test Final Order defines IMC as either the cost of an efficient device minus the cost of the standard device (ROB), or the full cost of the efficient device plus installation costs (simple retrofit). However, the Order also permits EDCs to utilize the Early Replacement calculation methodology, provided the EDC documents which method they used and why. The SWE incremental cost database remains an optional resource for EDCs and their ECs. EDCs may elect to use the cost assumptions in the incremental cost database or other reputable industry sources in their EE&C plans and annual TRC reporting. The source of all IMC assumptions should be documented. EDCs should use actual project costs where available and practicable (e.g., retrofit projects).

3.7.4 Data Requirements

To quantify the benefits of energy efficiency and evaluate the cost-effectiveness of individual measures, programs, and EE&C portfolios, evaluators must develop significant general modeling and measure/program-specific data assumptions. A full discussion of these data requirements can be found in the 2026 TRC Test Final Order.[80] Below is a brief list of these data requirements:

- General Modeling Assumptions
  - Avoided generation energy costs
  - Avoided generation capacity costs
  - Avoided transmission and distribution capacity costs
  - Avoided cost of compliance with Alternative Energy Portfolio Standards
  - Avoided costs due to price suppression effects
  - Avoided costs due to reductions in arrearages and collection costs
  - Energy and peak demand line losses
  - Discount rate (5% nominal) and general rate of inflation (2%)
- Program-/Measure-Specific Assumptions
  - Number of participants
  - Annual energy (kWh) savings
  - Annual summer demand (kW) savings
  - Annual winter demand (kW) savings
  - Annual water and fossil fuel impacts
  - Effective Useful Life
  - Incremental Measure Cost
  - Avoided Operations and Maintenance (O&M) benefits
  - Outside rebates/tax credits (if quantifiable)
  - Program administration (non-incentive) costs
  - Program/measure six-period load shapes

3.7.5 Cost Categories and Considerations

Program cost tracking should clearly delineate the categories needed for the cost-effectiveness assessment. Table 20 below lays out the cost categories and key considerations laid out in the 2026 TRC Test Final Order. The distinction between incentives and non-incentive spending is crucial given that incentives are not treated as a cost or a benefit under the TRC test and the Phase V Final Implementation Order established a limit on non-incentive spending at the portfolio level. In addition, there are overarching considerations for cost reporting:

---

[80] Pennsylvania Public Utility Commission, *2026 Total Resource Cost Test Final Order*, Docket No. M-2024-3048998, November 7, 2024.

- **Incremental cost assumptions**: For all measures, incremental cost (IMC) values must be clearly documented, and actual project costs should be used where available. EDCs may elect to use the cost assumptions in the SWE Incremental Cost Database or other reputable industry sources in their EE&C plans and annual TRC reporting. Note that reasonably quantifiable outside incentives, such as federal tax credits, are treated as a reduction in IMC.
- **Cases where incentives are greater than IMC**: Incentives must be reported at the measure level as an important input determining IMCs. Specifically, when the incentive amount is greater than the IMC, the incentive amount should be used as the TRC cost instead of the IMC.
- **Kit delivery as incentive and IMC**: As specified in the 2026 TRC Test Final Order, the cost of energy efficiency kits and directly installed equipment costs will be treated as IMCs and incentives.
- Inflation to 2026 dollars for phase-to-date reporting: As defined in the 2026 TRC Test Final Order, and consistent with prior practice, costs and avoided costs will continue to be provided in nominal dollars. A 2% inflation rate will be used to inflate values to 2026 dollars for the purposes of phase-to-date reporting across program years.
- **Non-incentive granularity**: costs may be tracked at the solution/component/sub-program level or be classified as cross-cutting cost assigned at the portfolio level.

Table 20: Cost Reporting Categories and Considerations

Cost Type	Cost Element	Definition and Considerations
Incentives	Rebates	Excludes direct install equipment costs and costs for EE&C kits.
	Upstream / Midstream Buydown	Financial incentive paid to manufacturers, retailers, or distributors to reduce the upfront cost of efficient equipment.
	Kits	Counted as IMC and incentive
	Direct Install Materials & Labor	For direct install measures only. Training and coaching costs for Strategic Energy Management and Retro Commissioning programs should also be included in this category.
Incremental costs	IMCs	Cost of efficient measure relative to baseline. Varies by measure vintage (see Table 19). When the incentive amount is greater than the IMC, the incentive amount should be used as the TRC cost instead of the IMC.
	Participant Cost net of Incentives	Out of pocket cost to the participant. Calculated as IMC minus incentives
	Incentives from Outside of Act 129	Any reasonably quantifiable non-Act 129 incentive, such as a rebate, tax credit, or grant, where the EDC has direct data on the amount of the incentive and the fact that the customer made use of the funds should be treated as a reduction in cost.
Non-Incentives	Program Design	Includes direct costs attributable to plan and advance the programs. For example, the design of a HER program should be included here, while the actual development and mailing of HERs would be attributable to Program Delivery
	Administrative	Includes rebate processing, tracking system, general administration, program management, general management and legal, and technical assistance. Any common portfolio costs that are allocated across programs fall in this category

Cost Type	Cost Element	Definition and Considerations
	EDC Program Delivery Cost	Direct program implementation labor and material costs incurred by the EDC.
	CSP Delivery Fees	Direct program implementation costs incurred by a conservation service provider (CSP) and invoiced to the EDC. This category includes labor, fuel, and vehicle operation costs for appliance recycling and direct install programs. For behavioral programs, this includes the printing and postage of HERs. If a CSP contract is structured on a pay for performance basis, those fees should also be included here.
	Marketing	Includes labor and materials incurred by the EDC, the marketing CSP, and implementation CSP to market and promote the program.
	EM&V	Includes fees from the EC and EDC labor and materials to support the EM&V process.
	SWE Audit Costs	Treated as a cost for the TRC test but excluded from the 2% spending cap for Act 129 EE&C programs.
	Other	Only included if necessary. Must be described.

3.7.6 Benefit Categories and Considerations

Table 21 lays out the benefit categories and any considerations described in the 2026 TRC Test Final Order. In addition, there are overarching considerations for benefit reporting:

Use of reported savings for quantifying impacts: when verified gross and verified net impacts are available or necessary they should be used to calculate TRC benefits. However, for established offerings or initiatives with relatively consistent savings performance, EDCs are encouraged to reduce evaluation costs by rotating evaluations so that every program does not get an impact evaluation every year. This will result in some unverified savings for programs that only have reported gross savings available for the annual TRC reporting. In these cases, reported savings should be used to perform TRC calculations.

Increased fuel consumption from fuel switching: should be considered as a negative TRC benefit.

Table 21: Benefit Reporting Categories and Considerations

Benefit Type	Benefit Element	Considerations
Avoided Cost of Supplying Electricity	Avoided Cost of Electricity	Time differentiated using the six costing periods established in the 2026 TRM and 2026 TRC Test Final Order (summer on-peak, summer off-peak, shoulder on-peak, shoulder off-peak, winter on-peak, and winter off-peak).
	Avoided Cost of Generation Capacity	Based on actual BRA clearing prices for years that were available at the time of EE&C Plan development. Values for the remaining years are forecasted according to the method outlined in the 2026 TRC Test Final Order and ACC.
	Avoided Cost of Transmission Capacity	Values are provided in the ACC or Table 1 of Exhibit 4 of the 2026 TRC Test Final Order.
	Avoided Cost of Distribution Capacity	Values are provided in the ACC or Table 1 of Exhibit 4 of the 2026 TRC Test Final Order. This benefit stream is not applied to participants who take service at primary voltage (e.g., Large C&I).

Benefit Type	Benefit Element	Considerations
	Compliance with AEPS	AEPS costs to be escalated using the 2% inflation rate over the forecast horizon. For consistency with the ACC and 2026 TRC Test Final Order, the AEPS avoided cost shall be $6.88 per MWh.
	Price Suppression Effects	Included as a benefit for the first time in Phase V. Individual values for each EDC by program year can be found in the ACC or Exhibit 3 of the 2026 TRC Test Final Order.
Other TRC Benefits	Water Impacts	Quantification: only required for measures where either the 2026 TRM provides all necessary inputs and assumptions to calculate them or the 2026 TRC Test Final Order presents default savings levels Monetization: $0.011 per gallon (in 2026 dollars) as the marginal cost of water used for TRC testing escalated annually over the forecast horizon, with a loss factor of 24.5% (1.32 multiplier) to be applied to all savings calculated at the premise level.
	Reduction in Arrearages and Collection Costs	Included as a benefit for the first time in Phase V. Individual values for each EDC by program year can be found in the ACC or Exhibit 2 of the 2026 TRC Test Final Order.
	Fossil Fuel Impacts	Quantification: required for fuel-switching measures, lighting interactive effects, and additional measure categories described in the 2026 TRC Test Final Order. Monetization: using natural gas avoided costs for Phase V specified in the ACC.
	O&M Benefits	Continue to include O&M benefits in the TRC Test as either positive or negative TRC benefits
	Societal Benefits	This benefit has not historically been included, and that practice will continue.

3.7.7 Annual Reporting Template

Section 3.7.5 describes distinct cost categories and their treatment in the context of determination of incentives and incremental costs for the purposes of TRC calculations. The table below presents a modification of the portfolio annual reporting table, which aligns with these categories.

Table 22: Summary of Portfolio Finances – Gross Verified

Row	Cost Category	PYTD ($1,000)		P4TD ($1,000)	
1	**IMCs**				
2	Rebates to Participants and Trade Allies				
3	Upstream / Midstream Incentives				
4	Material Cost for Self-Install Programs (EE&C Kits)				
5	Direct Installation Program Materials and Labor				
6	Incentives from Outside of Act 129				
7	Participant Costs (Row 1 minus the sum of Rows 2 through 6)				
8	EDC Labor Costs to Support Braided Funding				
		EDC	**CSP**	**EDC**	**CSP**
9	Program Design				
10	Administration and Management				
11	Marketing				

Row	Cost Category	PYTD ($1,000)	P4TD ($1,000)
12	Program Delivery		
13	EDC Evaluation Costs		
14	SWE Audit Costs		
15	**Program Overhead Costs (Sum of rows 9 through 14)**		
16	**Total NPV TRC Costs (Sum of rows 1 and 15 minus row 6 plus row 8)**		
17	Total NPV Lifetime Electric Energy Benefits		
18	Total NPV Lifetime Electric Capacity Benefits		
19	Total NPV Lifetime Operation and Maintenance (O&M) Benefits		
20	Total NPV Lifetime Fossil Fuel Impacts		
21	Total NPV Lifetime Water Impacts		
22	**Total NPV TRC Benefits (Sum of rows 17 through 21)**		
23	**TRC Benefit-Cost Ratio (Row 22 divided by Row 16)**		

3.8 Frequency of Evaluations

As mentioned in Section 3.5.1, every program (or initiative) should have at least one process evaluation in Phase V. Similarly, net impact evaluations need to be conducted at least once during the phase, but likely no more than three times. During the first three phases of Act 129, gross impact evaluations were typically completed annually. For Phase IV of Act 129, the SWE encouraged ECs to stage gross impact evaluations to encourage deeper investigations while managing EM&V expenditures and a compressed annual reporting timeline. Parties generally agreed that the flexibility to propose a rotating impact evaluation schedule across the five-year phase was a positive change in Phase IV. For Phase V, evaluators are encouraged to leverage multi-year samples or every-other-year impact evaluation designs, provided the reduced frequency translates to more robust research designs for the impact evaluations that are completed.

A rotating impact evaluation approach creates challenges given the annual reporting schedule. Specifically, how should MWh and MW savings from an initiative be reported if an impact evaluation was not completed during the program year? There are two possible approaches:

1. Present the energy and demand savings as unverified until the next impact evaluation is complete. Once the impact evaluation is complete, adjust all reported savings by the applicable realization rates.
2. Use a historic realization rate to adjust the reported savings in years when no new impact evaluation is completed.

Evaluators are expected to rely on a mixture of the two approaches for Phase V EM&V. In the case of a small and stable initiative, there is little risk associated with using a historic realization rate for a subset of the Phase V programs. Initiatives that operate in a rapidly changing market, or experience changes of ICSP, codes and standards, or measure mix are poor candidates for using a historic realization rate.

Historic realization rates from a prior phase are rarely appropriate and are especially problematic for Phase V because winter kW savings were not reported in Phase IV.

The EDCs should use the following criteria to propose the frequency and handling of reported savings during off years for every program or initiative:

- **Amount of energy and demand savings**. More frequent gross impact evaluations are warranted for programs or initiatives that are expected to produce the most energy and demand savings; conversely, programs and initiatives with low savings levels may not warrant annual gross savings evaluations.
- **Expected EM&V Costs**. Behavioral programs are one of the least expensive initiatives to evaluate because they rely on a straightforward analysis of billing data. While the expected savings contribution of HER initiatives in Phase V may not warrant an annual impact evaluation, the low cost of completion might encourage EDCs and their ECs to propose an annual cadence. An initiative with significant primary data collection requirements, in contrast, might make sense to evaluate two or three times in the phase.
- **Program continuity / discontinuity**. New initiatives and initiatives undergoing changes in measure composition, efficiency levels, incentives, program delivery, or implementation contractors likely warrant gross savings evaluations and possibly net savings evaluations and process evaluations within a year or two after those changes take place. In contrast, a program or initiative that remains largely unchanged, and with consistent realization rates year after year, could probably do with gross savings evaluations conducted every other year, and with net savings evaluations and process evaluations conducted only once in the cycle.
- **Market or technology continuity / discontinuity**. Changes in a market or to codes and standards may suggest more frequent evaluations or logical breakpoints for impact evaluations. Consider the residential heat pump water heater measure. Based on known code changes, the 2026 TRM calls for broad shift in baseline efficiencies at the beginning of PY21. EDC ECs might choose to leverage this change and conduct one impact evaluation prior to the code change and a second impact evaluation after the code change. This approach would avoid a set of realization rates calculated from a mix of standards.
- **Uniformity of measures**. If the efficient measures promoted by a program or initiative are the same year after year, then, other things being equal, it may not be necessary to evaluate that program every year. If the mix of measures varies from year to year, however – as with custom programs – then the savings would likely also vary, and more frequent gross impact and net impact evaluations would be justified.
- **Underperforming expectations**. If realization rates are disappointing and the evaluation leads to corrective actions, EDCs and their ECs may elect to increase the frequency of impact evaluations. Unexpectedly high realization rates can also indicate issues with the reported savings process and lead to program delivery modifications. Generally, the SWE suggests EDCs consider accelerating the planned impact evaluation schedule and limiting the application of historic realization rates when energy or demand realization rates are less than 80% or greater than 120%.
- **The reliability of ICSP savings claims**. For daily load shifting and behavioral programs, EDCs are not required to present reported gross savings in their semi-annual reports or

supply the SWE with ex ante values in their quarterly data request responses. As discussed above, HER programs have traditionally been evaluated annually because the cost is low and evaluators have established processes. The accuracy of ICSP estimates has varied by vendor, particularly with respect to persistence calculations and peak demand savings estimates. EDCs are free to present unverified savings for one or two years in a row and then analyze HER programs in a large multi-year analysis. If evaluation results show consistent alignment with ICSP estimates, the SWE would entertain proposals of a historic realization rate being applied to certain years of HER program activity. Many of the same considerations apply to daily load shifting. Evaluators should propose an annual evaluation for at least the first two years of Phase V, but a rotating focus could be considered for PY20 –PY22, conditional on a high degree of alignment between evaluated results and ICSP estimates in PY18 and PY19.

Each EDC should use the above criteria to propose preliminary five-year evaluation schedules for every program and initiative. The proposed schedules will be a central component of the SWE's EM&V plan reviews. The EDC EM&V plans should include the rationales for the schedule and off-year reporting method for each program and initiative. Table 23 shows a hypothetical impact evaluation overview table with two rows for each initiative. The first row indicates the sampling and data collection frequency, or which years of program activity each impact evaluation will examine. The second row shows how savings from the initiative will be presented in that year's final annual report, where:

- **V** = verified using the results of the impact evaluation completed that year.
- **H** = verified using the results of a historic impact evaluation.
- **U** = unverified until the results of the impact evaluation are available.

Table 23: Hypothetical Gross Impact Overview Table

Initiative	PY18	PY19	PY20	PY21	PY22
Offering #1 Sampling	Two-year sample		Two-Year Sample		None
Offering #1 Reporting	H	V	H	V	H
Offering #2 Sampling	Impact	None	Impact	None	Impact
Offering #2 Reporting	V	U	V	U	V
Offering #3 Sampling	Impact	Impact	Impact	Impact	Impact
Offering #3 Reporting	V	V	V	V	V
Offering #4 Sampling	Three-Year Sample			Two-Year Sample	
Offering #4 Reporting	H	U	V	H	V
Offering #5 Sampling	None	Impact	None	Two-Year Sample	
Offering #5 Reporting	U	V	H	H	V
Offering #6 Sampling	Impact	None	Impact	None	Impact
Offering #6 Reporting	V	H	V	H	V

The permutations shown in Table 23 are intended to be illustrative, not an exhaustive list of acceptable configurations. EDC ECs are encouraged to share draft tables of impact, NTG and process evaluation activities for SWE review prior to developing the full EM&V plan for Phase V.

3.9 Pilot Programs

The PUC's Phase V Final Implementation Order included multiple discussions of pilot programs and the Commission's expectations regarding the types of details that should be provided either in an initial EE&C Plan or in an EE&C Plan change. Key details include the pilot design, hypotheses being tested, budget, participation estimates, savings estimates, and timeline. Multiple stakeholders provided suggestions regarding potential Phase V pilots or the types of information they would like to see reported from pilots. Additionally, the PUC's Final Act 129 Phase V EE&C Plan Template[81] asked EDCs to confirm that no more than 2% of plan funds shall be allocated for experimental equipment or devices. Table 24 lists the pilots proposed by the EDCs in their Phase V EE&C Plans.

Table 24: Proposed EDC Pilots

EDC	Pilot	Description
PECO	Hybrid Heat Pump Load Shifting Pilot	This pilot focuses on customers with a hybrid heat pump system with both electric and gas heating. During winter peak demand periods, the electric heat pump is temporarily disabled, allowing the hybrid system to rely solely on the gas furnace for heating.
PECO	Low- and Moderate-Income Pilot	This pilot will identify, implement, and track comprehensive energy efficiency measures for LMI customers while supporting PECO's overall energy savings objectives. The pilot will test and refine incentive designs, marketing approaches, and enrollment processes that improve customer participation and installation of comprehensive measures.
PPL	Conservation Voltage Reduction Pilot	The objective of this pilot is to create and implement a methodology to reduce energy consumption and peak demand by reducing voltage in front of customer meters.
PPL	Peak Time Rebate (PTR) Pilot	This pilot tests whether a static, seasonal rebate with a performance bonus encourages residential customers to shift energy usage away from peak periods during both summer and winter.
PPL	Window Saddle Heat Pump Pilot for Resource Constrained Customers	This pilot will explore whether this nascent technology can effectively replace baseboard resistance heat in multifamily applications. Metrics to be tracked include product performance, customer response, and avoided electric demand and energy savings impacts.
PPL	C&I Peak Load Shift	This pilot will leverage existing building management systems, process load control, battery storage, thermal storage, on-site generation, or other methodologies to shift demand from peak hours to off-peak hours daily. The pilot will examine participant interest, customer satisfaction, software viability, and cost-effectiveness to inform potential expansion.

Given the limited budget allocation, pilot programs are not expected to contribute meaningfully toward Phase V compliance reduction targets. However, the primary goal of pilots is to assess the viability of innovative new technologies and program designs and determine whether they should be scaled up or abandoned. Answering this question should be the central focus of evaluation activities for pilot programs. Evaluators should design a plan capable of answering the core technical and economic

[81] *Final Act 129 Phase V EE&C Plan Template*, at page 25. Issued September 8, 2025, at Docket No. M-2025-3052826.

questions necessary to make an informed expansion decision and work in an embedded fashion with the EDC and its ICSP to carry out that plan. Like all impact evaluations, pilot evaluations must deal with counterfactual issues, data availability, sample size, and other fundamental issues. Evaluation elements that are unique or more pressing for pilots include:

- **Does the technology or intervention work as intended?** For the established measures in the TRM, there is little doubt that the technologies work as intended and deliver savings in a reasonably predictable manner. With pilot programs and nascent technologies, there could be fundamental technical, communication, customer acceptance, or market issues that limit the viability of the offering.
- **Measurement accuracy is more important than aggregate savings.** For most EE&C programs, production of energy and demand savings is the primary objective and savings quantification is an important secondary objective. Because pilot scopes are limited, but their findings are used to make far-reaching portfolio design decisions, producing correct inferences outweighs the claimable savings. This allows for more randomization and experimental design than a typical EE&C program. Typically, EM&V will account for a larger share of a pilot budget than a program budget.
- **Reduced applicability of TRM protocols.** Often, the goal of a pilot is to determine if a technology merits addition to the TRM and, if so, to develop a TRM measure characterization. Because the program design or technology is not as well understood, pilots will generally include primary data collection with a greater share of program participants.

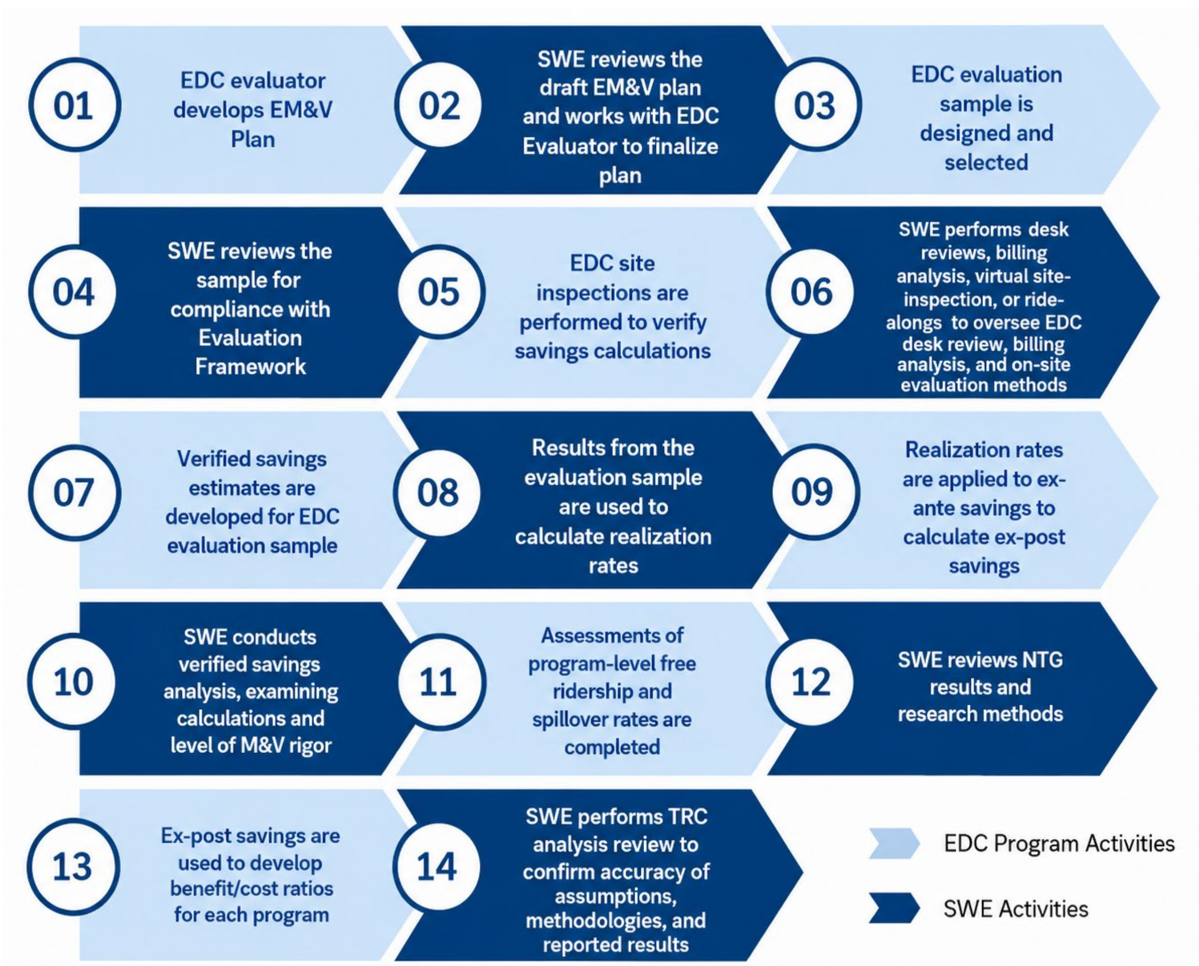
Pilot evaluation reports can be shared with the SWE and TUS either as an addendum to the EDC's Final Annual Report or as a standalone memo or presentation. At minimum, pilot evaluations should include:

- A detailed description of the pilot design and implementation, including the number of participants and the system used to track participation
- Evaluation methodology and data sources
- A description of the assumed baseline and how the pilot achieved improved energy outcomes relative to that baseline condition
- Energy and demand estimates with uncertainty
- A discussion of pilot economics and how cost-effectiveness might change at scale
- A summary of the strengths, weaknesses, opportunities, and threats of expanding the offering from a pilot to a full program
- Key findings and recommendations

4 Statewide Evaluation Audit Activities

This section describes the actions and activities conducted by the SWE to audit the implementation and the evaluation of each EDC's EE&C plan. This includes review/audit of EDC program delivery mechanisms and all evaluation processes and results submitted by each EDC's EC. The overall SWE audit findings should be used to inform the EDC evaluation teams when conducting the actual program evaluations. The SWE will use the audit activity findings, which will parallel the EDC evaluation activities, to assess the quality and validity of the EDC gross-verified savings estimates, net-verified savings estimates, process evaluation findings and recommendations, and benefit/cost ratios (BCRs). Figure 7 shows the specific SWE audit activities and their correspondence to the evaluation steps.

Figure 7: SWE Audit Activities



To the extent possible, the SWE will provide the EDCs with early feedback on the results of its audit activities – particularly if discrepancies are identified. The intent of early feedback is to allow the EDCs to work with ICSPs and ECs to implement corrective actions within the program year.

4.1 EDC Report and SWE Report Schedule

The semi-annual and final annual reports defined by the PUC are one of the ways in which stakeholders are informed about the spending and savings impacts of Act 129 EE&C plans. These semi-annual and final annual EDC and SWE reports are public documents. This section of the Framework provides an overview of the EDC and SWE reporting requirements for Phase V.

4.1.1 EDC Report Schedule

The EDCs are required to submit semi-annual and final annual reports to the SWE Team and the TUS. In the Phase V Final Implementation Order entered June 18, 2025, the PUC noted that Act 129 requires EDCs to submit a final annual report documenting the effectiveness of their EE&C plans, M&V of energy savings, evaluation of the cost-effectiveness of their expenditures, and any other information the PUC requires.

The SWE Team provides the EDCs with semi-annual and final annual report templates, which are available on the PA Act 129 SharePoint Site. The deadlines for the EDC reports are provided in Table 25.

Table 25: EDC Reporting Schedule

Report	Due	Savings Reported
Program Year X, Semi-Annual Report	January 15	▪ EE and PDR participation and impacts from Q1–Q2, including verified PDR for summer peak daily load shifting ▪ Implementation and evaluation updates ▪ Gross reported EE and PDR savings PYTD ▪ Sum of Incremental Annual Phase V savings (progress toward goals)
Program Year X, Final Annual Report	September 30	▪ Impact evaluation results (realization rates and confidence intervals) ▪ Gross verified EE and PDR savings (PYX) ▪ NTG results for measures and programs ▪ Process evaluation findings and recommendations ▪ TRC ratios at the program and portfolio level ▪ Sum of Incremental Annual Phase V savings (progress toward goals)

The semi-annual reports and final annual reports shall be filed with the PUC's Secretary and the SWE Team via the PA Act 129 SharePoint Site. The PUC will post these reports on its website for public access. The EDC final annual report template will also include a section requesting a comparison of actual program performance to the planning projections filed in their EE&C plans. Requested items will include the following:

- How did expenditures in the program year compare to the budget estimates set forth in the EE&C plan?
- How did program savings compare to the energy and peak demand savings estimates filed in the EE&C plan? Discuss programs that exceeded and fell short of projections and what factors may have contributed.

- Are there measures that exceeded or fell short of projected adoption levels? Discuss those measures, if any.
- How did the program year TRC ratios compare to the projected values in the EE&C plan?
- Are any changes to the EE&C plan being considered based on observations from the previous program year?

EDCs are required to correct errors that the SWE finds in their final annual reports to the Pennsylvania PUC in the following year's final annual reports if the change in annual portfolio kWh savings, kW savings, or TRC ratio is less than 5%. In instances where the change is greater than 5%, the EDC must correct such errors and refile the final annual report. All errors observed in the last final annual report of a Phase must be corrected and the Report must be refiled with the PUC.

4.1.2 Statewide Evaluator Report Schedule

In Phase V, the SWE Team will submit two reports to the PUC each program year. By February 28, the SWE will submit a semi-annual report summarizing and auditing the EDCs' semi-annual reports. By November 30, the SWE will submit a final annual report summarizing and auditing the EDCs' final annual reports. The final annual report will include the following information:

- Summarized program and portfolio achievements to date for each EDC
- Summarized energy (MWh/yr) savings and peak demand (MW) reductions for the program year and the sum of incremental annual savings progress toward the target for each EDC
- An analysis of each EDC's plan expenditures and an assessment of the program's expenditures
- An analysis of the cost-effectiveness of each EDC's expenditures in accordance with the Commission adopted TRC Order
- Identification of best practices exhibited to date
- Identification of areas for improvements
- An analysis of each EDC's protocol for M&V of energy savings attributable to its plan, in accordance with the Commission-adopted TRM, framework protocols, and approved custom measures
- A summary of SWE audit activities and findings based on the audit work completed

The reports also will include a summary of general activities corresponding to the responsibilities of the SWE Team. This could include the status of resolutions from any meetings/discussions and/or a summary of recently issued guidance memos.

The deadlines for the SWE reports to the PUC are presented in Table 26.

Table 26: SWE Reporting Schedule

Report	Due	Savings Reported
DRAFT Program Year X, Semi-Annual Report	February 14	▪ Summary of EDC-verified PDR impacts ▪ SWE PDR audit findings ▪ Summary of EDC-reported EE savings ▪ Summary of SWE Team EE audit activities and findings ▪ Draft report will be sent to the EDCs for review
FINAL Program Year X, Semi-Annual Report	February 28	▪ Final semi-annual report; comments from TUS staff and EDCs addressed
DRAFT Program Year X, Final Annual Report	October 27	▪ Summary of EDC gross verified savings claims from EE and PDR programs ▪ Review of EM&V practices and alignment with TRM and Evaluation Framework ▪ Summary of NTG and process findings ▪ Summary of SWE audit activities and findings ▪ SWE recommendations to accept or modify EDC savings claims toward statutory targets ▪ Summary of EDCs' sum of incremental annual savings toward targets
FINAL Program Year X, Final Annual Report	November 30	▪ Final annual report; comments from TUS staff and EDCs addressed

4.2 Reported Savings Audit

The SWE will conduct quarterly audits of the ex ante savings values claimed by EDCs and stored in EDC tracking systems. These audit activities are intended to give the PUC confidence in the gross reported savings values presented in EDC semi-annual and final annual reports. Gross reported savings estimates are the basis upon which the ex post evaluation is conducted.

4.2.1 Quarterly Data Request – Ex Ante

In a standing quarterly data request memo, the SWE Team requests information and data from the EDCs pertaining to the program implementation and the reported participation and savings associated with the implementation activity in the quarter. EDC quarterly data request responses are uploaded to the EDC-specific Act 129 SharePoint site and archived for various audit and reporting functions.

All information provided in response to the SWE data request should correspond to activities occurring during the quarter for which the EDC will claim savings. The sum of the kWh savings values in an EDC's data request responses for Q1 and Q2 should equal the PYTD kWh savings for that program in the EDC semi-annual report to the PUC. The SWE requires the following program-specific information for each program audit:

1. **Program Tracking Data** – A full export from the system of records listing the kWh, kW, rebate amount, participant information, and relevant dates for all transactions in the quarter. The SWE consolidates program tracking exports into a single statewide tracking database.

2. **Supporting Documentation** – For a subset of records in the program tracking data, EDCs are required to submit supporting documentation as defined in the SWE data request memo.[82] The type of supporting documentation varies by program delivery model but generally includes items such as application forms, equipment specification sheets, invoices for the purchase of efficient equipment, audit forms, and savings calculation workbooks (e.g., TRM Appendix C, D, F, or G).

4.2.1.1 Desk Audits

As part of its contract with the Pennsylvania PUC, the SWE will complete desk audits for all EDC programs. These audits will seek to verify the ex ante savings of EDCs' programs by collecting, recording, maintaining, and parsing EDC program data obtained via the SWE data requests described above. The SWE's desk audits will consist of the following three primary elements:

1. A **database review** through which the SWE will verify that EDCs are using the correct values and algorithms from the Pennsylvania TRM in their savings calculations. For deemed measures, the SWE will verify that the EDC used the correct deemed savings value unless otherwise approved by SWE and TUS. For partially deemed measures, the SWE will use the values from the EDC database to independently calculate savings and verify them against the savings reported by the EDC.
2. **Semi-annual and final annual report reviews** through which the SWE will verify that the values presented in EDC semi-annual and final annual reports match the values calculated by the SWE from the EDC database.
3. A **sample check** through which the SWE will cross-check actual program files, receipts, invoices, and work orders against their corresponding database entries to verify that the EDCs have reported program data correctly and consistently. This project file review is designed to audit the accuracy of the savings values stored in the EDC tracking system and to confirm that the EDCs' calculations were performed in accordance with the current TRM. The uploaded project files include project savings calculation workbooks, specification sheets for equipment installed, invoices, customer incentive agreements, and post-inspection forms. Through these reviews, the SWE will verify that savings values recorded in project files and the program tracking database are consistent.

4.3 Verified Savings Audit

The SWE will conduct an annual audit of the gross impact evaluation methodology and results for each program in an EDC portfolio and will summarize the findings and recommendations in the final annual report for the program year. The intent of the audit is to ensure that the savings are replicable, provide confidence in the gross verified program savings documented in the EDC final annual reports, and promote transparency in the evaluation process. Most of the SWE's findings and recommendations will be addressed prospectively in TRM updates, evaluation plans, and other M&V protocols used by the EDC ECs. Data gathered during the audit of an EDC program may be supplemented with best practice recommendations and techniques from other EDCs or national sources. The focus of the SWE's

[82] The SWE quarterly and annual data request memos are posted on the SWE Team SharePoint site.

prospective recommendations will be to enhance program delivery and cost-effectiveness and improve the accuracy of savings protocols used by the ICSPs and EDC ECs. The SWE will present the findings and recommendations from its annual audit activities in its final annual report for each program year. Unless errors are discovered or the SWE has significant concerns about the methodology used to calculate verified savings for an EDC program, the SWE will recommend that the PUC accept the verified savings provided in the final annual report.

If an EDC reports program savings using more than one calculation methodology, the SWE will offer its professional opinion regarding which method produces the most accurate representation of the program impacts in the SWE's final annual report. This situation typically arises when an EDC believes that a TRM algorithm or value does not accurately reflect the impact of a measure or the conditions in its service territory. In such cases, the EDC EC will present the savings impacts using both the TRM savings protocol and the "off-TRM" protocol that the EDC's EC believes is more appropriate for the measure. The SWE will review the savings protocol proposed by the EDC's EC and provide a recommendation that the PUC approve or reject the protocol. The SWE's recommendation should not be construed as PUC approval because the PUC has the ultimate authority to approve or reject the savings calculated using the proposed protocol.

4.3.1 Survey Instrument Review

Participant surveys are the most common form of data gathering used by EDC ECs to collect information about program populations because it is possible to generate a representative and large sample size at relatively low cost. Surveys can be conducted online, in person, via mail, or over the telephone. During Phase V, the ECs must submit draft survey instruments (in advance of survey implementation) that include process, NTG, or impact evaluation questions to the SWE for review prior to implementation. A question whose responses will be used as a parameter in a deemed or partially deemed algorithm is an impact evaluation question. Impact questions for a deemed measure typically involve a straightforward verification that the measure was installed as recorded in the program tracking system. Impact questions for a partially deemed measure could include the size, efficiency, fuel type, replacement protocol, or any other input that affects the savings estimate for the installed measure.

The SWE Team should be alerted via email by EDC ECs once survey instruments have been uploaded to the SWE Team SharePoint site for review. The survey instruments should be thoroughly checked by the EDC EC for skip patterns and skip logic before being submitted to the SWE for review. The SWE will provide comments and suggest possible revisions within five business days. Evaluators should correct any errors in the survey identified by the SWE, but otherwise, evaluators are not required to change the survey instruments based on the SWE's feedback, they should consider the guidance carefully. If the evaluators do not receive comments from the SWE within five business days, they can begin implementing the survey. The intent of the SWE review is to confirm that the survey instrument is designed according to industry best practices, that the impact questions will produce accurate and unbiased estimates of program impacts, and that the process questions are clear and provide useful information for the process evaluation. The following list includes some of the issues the SWE will consider as it reviews survey instruments:

- Are the skip patterns adequately delineated? Are there any combinations of responses that will lead to key questions being omitted from the battery?
- Are any of the survey questions leading or ambiguous? (Improperly worded questions can compromise the reliability of survey results.)
- Are there any missed opportunities? Are there important questions that are not included in the battery, or are follow-up questions needed to answer the research questions?

4.3.2 SWE Annual Data Request

EDCs must submit a response to the SWE's annual data request the same day as the submittal of the EDC's final annual report for a program year (September 30). This request focuses on the ex post savings analysis the EDC EC used to calculate gross and net verified savings. Responses should be uploaded to the EDC-specific directory of the SWE Team SharePoint site in a folder titled "PY_ Annual Data Request Responses." As noted above, the EDCs are welcome to submit findings and supporting materials early for review by the SWE. Any materials submitted by August 15 will be reviewed by the SWE by September 15 of each reporting year.

The following sections summarize the core components of the SWE annual data request. The SWE annual data request also includes requests for supporting documentation to facilitate the NTG, process, and low-income measures proportion audit.

4.3.2.1 Program Database Review Calculations

For some programs or initiatives, the gross impact evaluation involves line-by-line calculations of energy and demand savings for each record in the program tracking data. Evaluators often supplement the fields in the tracking data based on their independent verification activities, such as:

- Equipment capacity and efficiency characteristic lookups from an ENERGY STAR Qualified Products List, AHRI certifications, Design Lights Consortium Qualified Products Lists, or other industry certification resources
- Average values from survey efforts or other primary data collection. This might include the average ISR collected via an online survey or the proportion of participants who have electric domestic hot water rather than fossil fuel for a low-flow device measure. In this context, evaluators apply the sample means to all records in the population when estimating gross verified savings
- Estimated unit energy consumption (UEC) based on regression coefficients in the TRM or UMP applied to equipment characteristics for an appliance recycling program

While this type of analysis may rely on sampling to collect estimates of population means, the gross verified savings calculations are performed for a census of program participants rather than a sample. Savings calculations must be transparent and replicable for the SWE to approve gross verified savings claims. If the calculations are performed in Microsoft Excel, a functional workbook with active formulas should be submitted to the SWE. If the analysis is performed in statistical programming software, the evaluator is required to provide the syntax files along with all relevant input and output files.

4.3.2.2 Evaluation Sample Disposition

For each program or initiative, the EC should provide a table that contains the following type of information for each project in the completed evaluation sample. If sampling is done on a rolling basis, EDC ECs are encouraged to submit this information in advance of the formal due date. The number of evaluation groups will vary by EDC according to the design of the portfolio. The underlined terms below may be used as column headers in the table.

- Unique Identifier: This field should correspond to an identifier variable provided to the SWE for the project in the quarterly tracking data for the program. This may be a rebate number, project number, or enrollment ID.
- Stratum: If a stratified sample design is used, identify the stratum from which the sampled project came.
- Selection Type: When the sample was designed, identify whether this project was a primary sample or an alternate.
- Evaluation Activity: Identify the type of evaluation activity performed to develop verified savings estimates for this project (e.g., phone interview, online survey, desk review, site or virtual inspection, building simulation, or multiple methods).
- M&V Approach: Identify the approach used to calculate the verified savings for this project (e.g., simple verification, IPMVP Option A–D, or other appropriate methodology).
- Meters Deployed: Indicate whether any type of logging equipment was deployed at this site to collect information on key parameters in the savings calculations. (Yes/No)
- Verified kWh/year: Report the verified annual kWh/year savings for the project.
- Verified Summer kW: Report the verified summer peak kW savings for the project.
- Verified Winter kW: Report the verified winter peak kW savings for the project.
- Verified MMBtu: Report the natural gas savings or increased net natural gas consumption for fuel switching measures.

Evaluators should provide the following if available: supporting documentation showing the sample selection for each evaluation group and any error roll-up sheets that show the calculation of error ratio/C_v and achieved precision for the evaluation group.

4.3.2.3 Consumption Analysis

For programs that utilize a regression-based analysis of monthly, daily, or hourly energy consumption from utility revenue meters or connected device telemetry, evaluators should provide the statistical code and input data sets used to estimate savings. This applies to programs using an attempted census of participants. Evaluators should also provide a data dictionary defining the variables in the data set. For this type of initiative, the EDCs' final annual report should include the model specification and the relevant regression output, such as the following:

- Causal identification strategy (randomized control trial, matched comparison group of non-participants, past/future participants) and associated equivalence checks
- Number of observations used, number of missing values, and attrition to the analysis due to various data management filters
- ANOVA table with degrees of freedom, F-value, and p-value
- R-squared and adjusted R-squared values
- Parameter estimates for each of the independent variables, including the associated standard error, t-statistic, p-value, and confidence limits
- Residual plots or other model validation graphics
- Variance Inflation Factors or other tests for multicollinearity

Evaluation activities for daily load shifting programs will generally fall under this category within the annual data request. The SWE requests that evaluators provide the supporting documentation for summer season analyses for early review since verified savings for summer DLS programs are claimed in the EDC semi-annual reports due January 15 of each program year.

4.3.2.4 Evaluation Sample Audit

The SWE will select a sample of projects from each evaluation group provided in response to Section 4.3.2.2 and provide the EDC EC with a list of the unique identifiers for those projects. Within 15 days of receiving the list of unique identifiers, EDC evaluators must provide the evaluation documentation and findings for each project. The SWE Team will conduct a desk audit of these projects to confirm the reliability of the savings estimates. Section 4.3.4 provides additional regarding these SWE desk audits.

The documentation and findings to be supplied by the EC will vary depending on the evaluation approach used. These items should include:

- Site-specific M&V plans (SSMVPs)
- Completed site inspection reports
- Savings calculation worksheets
- Photos taken during the site inspection
- Building simulation model input and output files or spreadsheet models used to calculate verified savings
- Consumption data used for an Option C analysis
- Data files from end-use metering
- Survey responses

4.3.2.5 TRC Model Audit

The EC should submit an electronic version of or provide the SWE access to the model(s) used to calculate the TRC ratios for each EDC program in the EDC final annual report. The TRC model(s) should contain all inputs and outputs to the BCR. Key inputs the SWE will examine include the following:

- Discount rate
- Line loss factors
- Avoided costs of generation energy and capacity as well as T&D avoided costs
- IMCs
- Program administration costs
- Verified savings
- Effective useful life of measures or measure groups
- End-use load shapes or on-peak/off-peak ratios used in benefit calculations

4.3.3 Sample Design Review

The precision requirements for the gross impact evaluation of Act 129 programs were described in Section 3.6.1. The SWE will review the EDC EC's sampling approaches at three stages during program evaluation.

1. **EM&V Plan** – A thorough evaluation plan is an essential component of a successful evaluation. Sample design is one of many issues addressed in the EM&V plan for a program. The plan should outline who will be contacted, how many will be contacted, what type of evaluation activity will occur, and when the evaluation activity is expected to occur. During its review of EDC EM&V plans, the SWE will consider the proposed sampling plan and request revisions, if needed. It is important to note that the EM&V plan is assembled in advance of the program year, so the sample design must be flexible enough to adapt if program participation patterns differ from expectations.
2. **Quarter 3 of the Program Year** – Within a month of the close of Q3 (i.e., by March 31) for each program year, ECs should submit an updated sampling plan for each EDC impact and process evaluation scheduled for completion and reporting in that year's Final Annual Report. At that point in the program year, it is possible to estimate the final disposition of the program population for the year more precisely. The SWE will approve the EDC EC's sampling plan for the program year via telephone or email exchanges. If the SWE has concerns about the sample size, sample disposition, or level of rigor used within the sample, the SWE will suggest modifications.
3. **SWE Final Annual Report** – Following the close of each program year, the SWE will review the evaluated results of each EDC program and provide recommendations for future program years. If the SWE feels a particular technology was under-represented in the evaluation sample, the SWE final annual report will contain a recommendation to focus more heavily on that technology or delivery mechanism in the next impact or process evaluation. If the evaluator's variability estimates (C_v or error ratio) proved to be too high or too low, the SWE will recommend changes to the sample design for the following year. As described in Section 3.6.1, impact evaluations that fail to meet the minimum precision requirements are not permitted to be used as historic realization rates. For programs that rely on participant surveys, the SWE will examine the sample frame carefully to assess whether there is any appearance of non-response bias or self-selection. If the SWE identifies any concerns, it will discuss the issue and suggest possible corrective actions.

4.3.4 Project Audits

Project inspections are essential for the accurate evaluation of programs and will represent a significant portion of the EDCs' evaluation efforts for programs. To ensure the accuracy and veracity of the EDC evaluation project inspections, the SWE will audit the M&V methods used by the evaluator to ensure the verified savings are calculated using approved protocols. To accomplish this, the SWE will request verification data annually for projects in the sample drawn by the EC for each EDC program. Typically, projects for the SWE Evaluation Sample Audit will be selected from the evaluation sample submitted as part of the SWE Annual Data Request. If an EC completes a significant share of the verified savings analysis for a program year in advance of the reporting deadline (September 30), the SWE will consider a multi-stage sampling process to allow increased discussion prior to the inclusion of audit findings in the SWE Final Annual Report.

SWE project audits will fall into three categories:

1. **Desk Reviews**: The SWE will annually review a sample of EDC evaluation project analysis findings and recommendations, as well as actions taken by EDCs to address them.
2. **Ride-Along Site Inspections**: The SWE may perform ride-along audits, for a small share of evaluated projects in which the SWE accompanies the EC on a site inspection to validate and confirm that ECs are using approved protocols when performing evaluation activities. This includes checking for adherence with the TRM, compliance with this Evaluation Framework, and use of industry best practices. The ride-along audits are a sub-set of the EDC evaluation sample, focusing on high-impact and high-uncertainty projects. The site-specific savings should be adjusted based on the SWE's findings and recommendations. The SWE expects to conduct more site inspection audits for new programs, new technologies, or new ICSPs. As the SWE becomes more confident in the accuracy of reported and verified estimates for high-impact and high uncertainty projects, fewer ride-along site inspections will be conducted.
3. **Independent Site Inspections**: Although much less frequent than ride-along audits, the SWE reserves the right to perform an independent audit of any project in the program population with either high impact or high uncertainty, as determined by the SWE at any point in the program year. This may include sub-samples of the EDC evaluation sample or projects outside the EDC evaluation sample. Independent site inspections will include a detailed assessment of the measures beyond what would be performed by the SWE during ride-along inspections, to ensure that the measures are being operated to yield the energy and demand savings claimed in the rebate application. As appropriate, independent site inspections will include spot measurements or trending of important performance parameters and independent verified estimates for energy and peak demand savings.

The SWE's report content and format will be nearly identical for desk review reports and ride-along site inspection reports, while independent site inspection reports (I-SIR) will differ. Below is a description of each report type. Findings from all three project audit types will be summarized in the corresponding SWE Annual Report. The SWE will produce and distribute its desk review reports and ride-along site inspection reports to EDC evaluators within 30 business days of receiving the verified savings calculations.

Desk Review and Ride-Along Site Inspection Reports: Reports will include a description of the project, a summary of the EC's methodology and findings, and a discussion of the SWE's review of the EC's methodology and findings. The SWE's review will encompass many aspects of the M&V and ensuing analysis the EC conducted, including proper TRM protocol application, data collection practices, and savings analysis method selection. The SWE will either confirm the EC's verified savings or recommend changes to the verified changes based on findings in the SWE's review.

The reports will be completed after the EDC evaluators have shared their site inspection reports and engineering calculations with the SWE. ECs will have the opportunity to review the SWE findings and discuss key issues and/or discrepancies with the SWE. Resolutions will be reached collaboratively by the SWE and the ECs.

Independent Site Inspection Reports: Reports will include process findings related to program delivery and an independent SWE verified savings estimate. Because independent site inspections are conducted on sites not selected by the EDC ECs, independent site inspection reports will be issued shortly after SWE evaluation activities have been completed.

If the SWE elects to conduct an independent site inspection, the EDC and EC will be notified in advance of the visit. Verified savings estimates from projects receiving a SWE independent site inspection report can be included in the gross impact evaluation sample and subsequent realization rate calculation at the discretion of the EC. ECs will not be required to incorporate the results from independent site inspection reports in the final realization rate calculations.

4.4 Net Impact Evaluation Audit

Any Act 129 net impact research will be audited by the SWE. Further, EDCs are expected to conduct net impact research to inform program planning.

4.4.1 Research Design

The SWE will audit the research design as part of the review of the EM&V plan, and again as part of the review of the reported results. The audit will assess whether the approach used is consistent with common methods recommended for downstream programs and for ARPs (Appendix B and Appendix C).

For programs that cannot use the common method, the audit review will be based on the SWE-defined levels of rigor of analysis summarized in Table 27.

Table 27: Rigor Levels Adapted from the California Evaluation Protocols

Rigor Level	Methods of Net Impact Evaluation (Free Ridership and Spillover)
Basic	Deemed/stipulated NTG ratio Participant self-reporting surveys Expert judgment
Standard	Billing analysis of participants and non-participants Enhanced self-report method using other data sources relevant to the decision to install or adopt a measure. These could include record/business policy and paper review; examination of other, similar decisions; interviews with multiple actors and end users; interviews with midstream and upstream market actors; and interviews with program delivery staff. Market sales data analysis Other econometric or market-based studies
Enhanced	Triangulation. This typically involves using multiple methods from the standard and basic levels, including an analysis and justification of how the results were combined.

Method selection should follow the recommended threshold guideline based on a program's contribution to total portfolio savings. If the energy savings of an EDC's program is less than or equal to 5% of the EDC's total portfolio energy savings, a basic level of rigor analysis (e.g., stipulated/deemed or simple survey) is acceptable to estimate NTGRs. If the energy savings of an EDC's program is greater than 5%, the SWE recommends a more complex approach to determine whether the basic, standard, or enhanced level of rigor were appropriate. These recommendations are based on benefit/cost considerations, as the added costs of a greater level of rigor generally are unwarranted for programs with low savings contributions.

4.4.2 Sample Design

The audit will determine whether the sampling was appropriate. Probability sampling (described in Section 3.6 and Section 4.5.2) should be used for net savings or market share/market effects studies. The sample design will be audited as part of the review of the EM&V plan, and again as part of the review of the reported results.

4.4.3 Transparency in Reporting

The audit requires that the EDC and their ECs describe the reasons the approach was selected, the sample, the questions used, and the methods used in the analysis and application of the NTGR. Such information should include methodology, data collection, sampling, survey design, algorithm design, and analysis. Free ridership or NTG ratios should include explanation or description regarding how they were derived. A transparent approach to net savings is necessary for an effective and useful audit.

4.4.4 Use of Results

The audit also will examine how the EDC and its ECs are using the results for the purposes of modifying and improving program design and implementation while operating within Act 129 budget, cost-effectiveness, and compliance constraints.

4.5 Process Evaluation Audit

The SWE will audit process and market evaluation research plans, data collection instruments, and final reports to ensure the following:

- Research objectives are complete, appropriate, and likely to lead to actionable findings relative to the type of process or market evaluation planned.
- The sample design is sufficient and appropriate to address the objectives.
- The data collection approaches are appropriate and executed per plan.
- The data collection instruments address the objectives and do not introduce bias.
- The analysis and report writing convey the findings clearly and draw reasonable conclusions.
- The recommendations are actionable and clearly identify which parties should address the recommendations.
- EDCs follow up on process evaluation recommendations and report to the SWE the actions the EDC has taken on each recommendation.

4.5.1 Guidance on Research Objectives

The SWE audit will review the process evaluation with expectations that the process evaluation will address objectives as appropriate to the program. Examples of objectives that may be relevant to a program are noted below.

4.5.1.1 Program Design

- Program design, design characteristics, and design process
- Program mission, vision, and goal setting and process
- Assessment or development of program and market operations theories and supportive logic models, theory assumptions, and key theory relationships - especially their causal relationships
- Use of new practices or best practices

4.5.1.2 Program Administration

- Program oversight and improvement process
- Program staffing allocation and requirements
- Management and staff skills and training needs
- Program information and information support systems
- Organizational barriers to program administration
- Reporting and the relationship between effective tracking and management, including both operational and financial management

4.5.1.3 Program Implementation and Delivery

- Description and assessment of the program implementation and delivery process
- Clarity and effectiveness of internal staff communications
- Quality control methods and operational issues
- Program management and operational practices
- Program delivery systems, components, and implementation practices
- Program targeting, marketing, and outreach efforts
- Available and needed resources for effective program implementation
- The level of financial incentives for program participants
- Program goal attainment and goal-associated implementation processes and results
- Program timing, timelines, and time-sensitive accomplishments
- Quality-control procedures and processes

4.5.1.4 End-User and Market Response

- Customer interaction and satisfaction (both overall satisfaction and satisfaction with key program components, including satisfaction with key customer-product-provider relationships and support services)
- Customer or participant energy efficiency or load reduction needs and the ability of the program to provide for those needs
- Trade allies' interactions and satisfaction
- Low participation rates or associated energy savings
- Trade allies' needs and the ability of the program to provide for those needs
- Reasons for overly high free riders or low levels of market effects, free-drivers, or spillovers
- Intended or unanticipated market effects

4.5.2 Sample Design

Sampling for process and market evaluations should follow sampling approaches like those used for impact evaluations whenever it is important to generalize to the population. (Note, this does not mean that the sampling should be the same for impact and process and market evaluation, just that the approaches when generalization is important are similar). Table 28 outlines the three primary options for sampling; all may be used with process and market evaluations when appropriate. Section 3.6.2 provides additional guidance on probability sampling.

Table 28: Sampling Options

Option	What Is Measured	Applicability of Precision Estimates	Rank Order of Defensibility
Census	Measures the entire population, so results represent the entire population	Statistical precision is not applicable because the analysis counts every outcome and, therefore, provides a full rather than partial enumeration.	Highest
Probability Sample: Simple random and stratified random	Measures a randomly selected subset of the population, therefore the probability of selection to the sample is known and results can be generalized to the population	Sampling precision depends on the number of items; e.g., participants measured. The more measured, the better the precision.	Varies
Systematic Sample: Any non-random method of sampling	Measures a non-randomly selected subset of the population, so the probability of selection to the sample is unknown, and generalization to the population is not possible	Statistical precision is not applicable. Carefully selected representative samples sometimes are claimed to have properties similar to probability samples.	Lowest

Non-probability samples sometimes are acceptable for process and market evaluations. When sampling from small groups in which a census or near-census is possible, precision and confidence do not apply, and a census or near-census should be pursued. Non-probability samples also are acceptable when the purpose is to gain a greater sense of knowledge of the topic and not to generalize. In such cases, systematic sampling is acceptable. Evaluators must ensure that they have used robust, systematic sampling approaches and have articulated the justification for using a non-probability sample clearly in the process evaluation section of the EDC final annual report.

The process and market evaluators must identify the population, prepare an appropriate sampling frame, draw the sample consistent with the frame, and ensure that inference is consistent with the sampling approach.

4.5.3 Data Collection Instruments

The SWE must review all data collection instruments (in advance of survey implementation) and will strive to complete the review within five business days per the guidelines below.

4.5.3.1 General Instrument Characteristics

The SWE reviewers will audit the instruments, scrutinizing various elements, as described below:

- Title: including contact type (e.g., program staff, participants, non-participants, trade allies, industry experts)
- Statement of purpose (brief summary for interviewer, client, and survey house)
- Listing and explanation of variables to be piped into the survey and the source of these values (if applicable)

- Instructions to the interviewer/survey house/programmer regarding how to handle multiple response questions (e.g., process as binary)
- Scheduling script: collect time and date for re-contact, verification of best and alternative phone numbers
- Brief introduction: mentions client and requests client feedback for appropriate purposes
- Statement as to whether responses will be treated as confidential or will not be reported
- Screening questions: if needed, and if interviewer instructions include directions regarding when to terminate the survey
- General flow: from general questions directed to all contacts through specific topics (with headings), including skip patterns where needed
- Insertion of intermittent text or prompts to be read by the interviewer, informing the contact of new topics that also serve to improve the flow of the interview
- Use of a standard set of demographic/firmographic questions (e.g., comparable to Census or industry data)
- If needed, request permission to call back or email with follow-up questions (especially useful when conducting in-depth interviews); collection of appropriate callback information, best phone, email address, etc.
- Request for any additional comments from the respondent
- Conclusion, with a thank-you message

4.5.3.2 Question Review

The SWE will check for and comment on questions that are:

- Double-barreled (this and that)
- Leading and/or biased (questions that encourage participants to respond to the question in a certain way)
- Confusing or wordy (editing for clarity)
- Appear not to be related to research issues or analysis plan
- Related to research issues or analysis plan but do not appear to achieve the research objectives
- Clearly indicate whether to read or not read responses and when multiple responses are accepted
- Missing a timeframe anchor (e.g., in the past year)
- Driven by a skip pattern (Survey developers and reviewers must check that the skip is necessary and applies to all contacts, if at all applicable. It is best to avoid skips within skips that reduce the size of the sample.)
- General readability

4.5.4 Analysis Methods

The EDCs must use the appropriate levels of analysis for process evaluation data. Inference from the data should be consistent with the sampling strategy, and claims should not overreach the data. Data will be either qualitative or quantitative.

4.5.4.1 Qualitative Analysis

The EDC evaluators should respect the respondents' rights and not report names or affiliations except at a general level (e.g., program staff, implementers, customers, contractors, and trade allies). Reports should clearly document the program activities and lessons learned from the research. Findings should permit the reviewer to understand the data source(s) for the finding and to understand how different audiences responded to the research objectives. The population always should be clearly defined, and all tables and reported data should clearly articulate the portion of the sample responding for the finding [e.g., 7 of 10 people, or seven said (n=10)] and that tables are clearly labeled.

4.5.4.2 Quantitative Analysis

The EDC evaluators should ensure that response dispositions are tracked and reported consistent with the guidance of the American Association for Public Opinion Research.[83] The population always should be clearly defined, and all tables and reported data should clearly articulate the portion of the sample responding for the finding [e.g., 70% (n=349)] and ensure that tables are clearly labeled.

Further, the EDC EC should use appropriate quantitative methods. For instance, if data are ordinal – means should not be used – the top two boxes are acceptable. If data are not normally distributed, non-parametric tests should be used. Similarly, evaluators should choose statistical tests and analysis methods carefully to ensure that they are appropriate for the data collection process.

4.5.5 Assessment and Reporting by the SWE

The SWE process evaluation assessment will include a review of findings and recommendations relative to program design, program delivery, administrative activities, and market response. The SWE may conduct the following additional reviews and summaries of process findings during Phase V:

- Identify best practices across the state.
- Compare process evaluation findings to process and delivery strategies of similar best programs throughout the United States.
- Highlight areas of success within the portfolio of EDC projects and that identifies areas of improvement.
- Report on selected EDC responses to the recommendations.

[83] See https://aapor.org/response-rates/ and https://aapor.org/wp-content/uploads/2023/05/Standards-Definitions-10th-edition.pdf .

4.6 Cost-Effectiveness Evaluation Audit

The SWE cost-effectiveness assessment will include a review of the benefit-cost ratio formulas, benefits, costs, and TRC ratios at the program, sector, and portfolio level. The SWE will determine whether TRC calculations have been performed according to the 2026 TRC Test Final Order and whether EDCs are on track to meet the Act 129 cost-effectiveness requirements.

4.6.1 Annual Data Request

The SWE Team will request each EDC to submit an electronic version of the model(s) used to calculate the TRC ratios in the EDC's final annual report. The TRC model(s) should contain all relevant general modeling and program-specific inputs to the benefit-cost ratio, calculation formulas, and TRC outputs, as well as the completed ACC.

4.6.2 Inputs and Assumptions

Key inputs and assumptions the SWE Team will examine include the following:

- Line loss factors
- Avoided costs of energy and capacity
- IMCs
- Program administration costs
- Verified savings figures
- Effective useful life of measures or measure groups
- End-use load shapes or on-peak/off-peak ratios used in benefit calculations

4.6.3 Calculations

Possible audit activities pertaining to the cost-effectiveness protocols, calculations, and evaluations may include, but are not limited to, the following:

- A review of TRC Order compliance regarding:
  - Formulas
  - Benefits
  - Costs
  - Utility avoided costs assumptions
- A review of EDC accounting practices, including the following:
  - Division of costs and benefits between programs
  - Appreciation/depreciation rates

For Phase V, EDCs may choose to adopt a proprietary benefit-cost software product for their TRC analysis. For EDCs using proprietary products, the SWE Team will perform, at a minimum, a thorough one-time benchmarking of the TRC calculations to verify that results are reasonable and accurate. EDCs would continue to be required to provide inputs and outputs to the SWE for annual reporting purposes.

4.6.4 Additional Activities

In addition to the detailed audits of TRC calculations and results for each EDC, the Phase V SWE will compare results across the four EDCs. This will enable the SWE Team to accomplish the following:

- Report on trends in results over time and across EDCs;
- Identify and investigate large directional differences between EDCs; and
- Cross check assumptions and support efforts to achieve consistency across EDCs for topics including but not limited to incremental costs and dual baselines.

The SWE will investigate and provide guidance on additional benefit-cost considerations as directed by the PUC. Further, the SWE will address key topics as directed in the 2026 TRC Test Final Order. These topics and key considerations are described in Table 29 below.

Table 29: Additional Audit Activities

2026 TRC Test Final Order Topic	Audit Activity Description	Considerations
Vintage of Avoided Cost Forecasts	Compare forecasted avoided costs of electricity to load-weighted real-time LMPs for each EDC service area	The Phase IV comparison revealed mixed alignment between forecasts and actual costs. Early in the phase, energy costs were higher than expected and capacity auction prices were low. This trend then reversed toward the end of the phase. Continuing comparisons going forward will inform methodology standardization and ACC development
Compliance with AEPS	Summarize the AEPS costs in the Phase V SWE final annual reports and identify any significant differences between the assumed forecasted AEPS and the actual future AEPS costs	Significant differences between the assumed forecasted AEPS and the actual future AEPS costs may trigger a mid-phase update to avoided cost forecasts

5 Resources and Meetings

This Evaluation Framework is intended to serve as a resource for EDC program administrators and evaluators. The Framework is a living document and may be updated periodically in Phase V; however, we suggest that stakeholders familiarize themselves with several additional resources to stay informed of the latest developments related to the evaluation of Act 129 EE&C plans.

5.1 Pennsylvania Act 129 Public Utility Commission Website

The SWE Team will provide documents for sharing on the PUC's public website,[84] which provides information to interested stakeholders on the actual kWh/year and kW savings from the Act 129 programs, as well as the EDCs' expenditures on such programs.

5.2 Pennsylvania Act 129 SharePoint Site

The SWE Team maintains an Act 129 SharePoint site to improve communication and coordination of activities among the SWE Team, the TUS, the EDCs and their ECs, and the Energy Association for Phase V. This SharePoint site serves as a repository of documents and data associated with the statewide evaluation of the EE&C Program Portfolios implemented by the four EDCs. The structure and operation of this SharePoint site comply with the confidentiality provisions in the SWE Team contract with the PUC and the Energy Association.

An individual SharePoint site is set up for each EDC, along with a common SharePoint site to share statewide documents and information applicable to all EDCs. Individual access to each site and pages within the site is based upon assigned administrator privileges and confidentiality of content and the Nondisclosure Agreement signed by all parties and referenced in the document "Contract Act 129 Statewide Evaluator" (Issuing Office: Pennsylvania Public Utility Commission, Bureau of Technical Utility Services; RFP-2025-2).

The PA Act 129 SharePoint includes the following:

- **Common SWE site** that provides a common interface for all parties directly involved in the statewide evaluation efforts and that have been granted access to the Act 129 SharePoint Site. This site includes guidance memos, report templates, data requests, approved IMPs, TRM codes and standards memos, SWE deliverables, and meeting notes from meetings between the SWE, EDCs, and Energy Association of Pennsylvania (EAP).
- **Internal SWE-TUS team site**, whose access is restricted to members of the SWE team and the TUS staff. The purposes of the SWE Team directory are to facilitate coordination of SWE Team activities, collaborate on various work products, track progress, and store lists of unresolved issues.

84 The URL for the Act 129 directory of the PUC's website. Website

- **Individual EDC password-protected sites**, which are tailored to each EDC's needs and include features such as submissions library, task lists, and memo libraries.

For Phase V, the SWE Team will create Level 1 folders to house documents such as reports, tracking data, and data requests/responses and Level 2 folders for each program year. The Level 1 and 2 folder structure will be consistent across the individual EDC sites.

5.3 Program Evaluation Meetings

The SWE Team will chair and set the agenda for as-needed meetings involving TUS staff, the EDCs, EAP, and EDC evaluators. The SWE Team will prepare minutes of these meetings. These meetings will be conducted in person or virtually if necessary. The SWE will prepare PowerPoint presentations as needed.

5.4 EDC-Specific Quarterly Meetings

The SWE Team will coordinate quarterly meetings with each EDC EC to discuss EDC-specific technical issues. EDC personnel typically attend these meetings, but TUS Staff and EAP do not, as these meetings focus on detailed technical matters rather than policy discussions.

5.5 Stakeholder Meetings

Key members of the SWE Team will attend stakeholder meetings and deliver presentations on the results of baseline studies, market potential studies, TRM updates, and recommendations for program modifications and targets for Phase VI of Act 129.

6 Measure-Specific Evaluation Protocols (MEPs)

6.1 Behavioral Conservation Programs Evaluation Protocols

Behavioral conservation programs, such as HER and Business Energy Report (BER), encourage conservation through greater awareness of consumption patterns and engagement with EDC resources to help reduce usage and lower bills. Behavioral program vendors provide participants with account-specific information that allows customers to view various aspects of their energy use over time. Behavioral reports compare the energy use of recipient homes and businesses with clusters of similar homes and businesses and provide comparisons with other efficient and average homes. This so-called neighbor comparison is believed to create cognitive dissonance in participants and spur them to modify their behavior to be more efficient. Reports also include a variety of seasonally appropriate energy-saving tips that are tailored for the home or business and are often used to promote other EDC program offerings. Historically, behavioral reports have been largely issued on paper via the USPS, but EDCs and their vendors are increasingly moving toward email reports and digital portals to promote increased engagement and conserve resources.

There are a growing number of behavior-based programs that EDCs may wish to consider in their EE&C plans. This protocol does not attempt to address all possible variants of behavior-based programs as the EM&V approach will necessarily vary widely depending on the program delivery strategy. Instead, it focuses on providing clear guidelines for claiming compliance savings for the most prevalent behavior-based programs in the Phase V EE&C plans approved by the PUC. The guidance in this protocol is largely applicable BER programs if an EDC elects to offer BERs in Phase V. Where applicable, the protocol addresses unique considerations for opt-in web portals that customers can enroll in and visit to engage with their energy consumption data and receive conservation tips. If EDCs choose to offer additional behavior-based programs, the proposed EM&V approach should be described in an EM&V plan and submitted to the SWE for review and approval.

6.1.1 Impact Evaluation

The objective of the impact evaluation is to estimate the verified energy (kWh) and peak demand (kW) impacts of the program. Energy savings are used to report progress toward Act 129 consumption reduction goals. Peak demand impacts are used to report progress toward Act 129 PDR goals. Both types of resource savings are needed when calculating benefits for the TRC test.

6.1.1.1 Experimental Design

Act 129 HER and BER programs should be implemented as either a randomized control trial (RCT) or randomized encouragement design (RED) to ensure the accurate and unbiased estimation of program impacts. An RCT is an experimental design in which eligible participants are randomly placed into either a treatment group or a control group. Only the treatment group receives the reports. Typically, behavioral programs are delivered on an opt-out basis, meaning that the program automatically enrolls participants (instead of the participant signing up) and will send treatment group households or businesses reports unless the participant formally indicates that they want to leave the program. An RCT is generally considered to be the gold standard of evaluation protocols because the randomization process ensures that the energy reports are the only plausible explanation for the observed energy savings, as long as the treatment and control groups used electricity in a nearly identical manner prior to the receipt of EDC energy reports.

An RED is a variant of the RCT design that allows for an opt-in program delivery model. In an RED, participants are randomly assigned to either a treatment or control group. However, instead of automatically receiving the intervention, treatment group participants are only encouraged to take part in the EDC offering. Web portals are an example of a behavioral offering that would be well-suited to an RED approach because only a subset of the homes encouraged to visit the web portal will do so.

The SWE's review of Phase V EE&C Plans did not reveal any behavioral offerings where randomization into treatment and control groups would be problematic, but new strategies are likely to emerge during a five-year phase. Any departure from an RCT (or RED) design for behavior-based offerings should be vetted with the SWE prior to implementation. When randomization is done correctly, impact estimation for behavioral programs is straightforward. The RCT design also eliminates the need for NTG analysis because the control group does everything the treatment group would have done. Although the estimated savings are technically net savings, EDCs should claim the measured behavioral impacts toward Act 129 gross verified compliance reduction requirements.

Random assignment to the treatment or control group is slightly more complex for BER programs because the definition of a customer is less clear-cut. For example, a single business account in the EDC billing system may be associated with multiple meters or premises. Having one meter or premise in the control group and the other in the treatment group could create customer confusion and potentially compromise the control group (if the BER caused the customer to conserve energy in both spaces). EDCs should work closely with vendors and ECs to develop a randomization strategy that makes sense based on the account/premise/meter distinctions in the billing system and preserves the integrity of the RCT.

Group Sizes

The absolute precision of behavioral impact estimates is a function of two factors:

1. Unexplained variability in customer electricity usage
2. The number of homes or businesses in the treatment and control groups

The magnitude of the treatment effect is only a factor when relative precision is considered. EDCs have little control over the first factor – and cannot know the size of the treatment effect in advance – so treatment and control group size are the real levers that the EDCs have to work with. When group sizes differ, the smaller of the two groups becomes the primary determinant of precision. Since participants in the control group produce no savings, the common approach is to make the treatment group larger than the control group.

As a result, the practical question related to precision is "How precise do the measurements of behavioral program savings need to be?" and, in turn, "How large do group sizes need to be to meet this precision requirement?"

- **For HER programs**, EDCs should design group sizes to produce an expected program-level absolute precision of ± 0.5% at the 95% confidence level (two-tailed) at the onset of treatment. Individual cohorts within an HER implementation may have a wider margin of error.
- **For BER programs**, EDCs should design group sizes to produce an expected program-level absolute precision of ± 0.5% at the 85% confidence level (two-tailed) at the onset of treatment. Individual cohorts within a BER implementation may have a wider margin of error.

The intent of this requirement is to ensure that HER and BER programs, which represent a sizable share of Phase V EE&C budgets and projected savings, are measured in a manner that makes the savings claims unassailable and supports an accurate assessment of whether the investment of rate-payer funds in this brand of energy efficiency is cost-effective. The SWE will review and approve on a case-by-case basis less precise designs for behavioral programs offered to targeted populations or populations of limited size where the ± 0.5% absolute precision is difficult or impossible to attain. Exceptions will also be considered for pilot offerings where EDCs wish to explore the effects of a new behavioral offering with a few thousand customers instead of committing limited resources to treat the tens of thousands of participants needed to achieve ± 0.5% absolute precision.

The ± 0.5% absolute precision requirement expresses the required margin of error as a function of annual consumption, not savings impact. If the average consumption for a household in an EDC HER program is 12,000 kWh per year, the program design should enable energy savings determination to within ± 60 kWh at the 95% confidence level. In a BER program where businesses use 40,000 kWh per year on average, this requirement would translate to an absolute margin of error of at least ± 200 kWh.

It is important to note that this requirement for program design is different from the sampling requirement, set forth in Table 16, that programs annually achieve ±15% relative precision at the 85% confidence level. Standard industry precision requirements are not reasonable expectations for behavioral programs because the size of the average effect is typically much smaller, and all estimation error is captured as opposed to sampling error only, like in most other programs.

Consider the residential example above where homes use, on average, 12,000 kWh annually and the HER program is required to produce impact estimates within ± 60 kWh at the 95% confidence level (± 44 kWh at the 85% confidence level). If the average treatment effect (ATE) in this example was 150 kWh per household annually, the relative precision at the 85% confidence level would be:

$$Relative\ Precision = \frac{Margin\ of\ error}{Average\ treatment\ effect} = \frac{44}{150} = 29.3\%$$

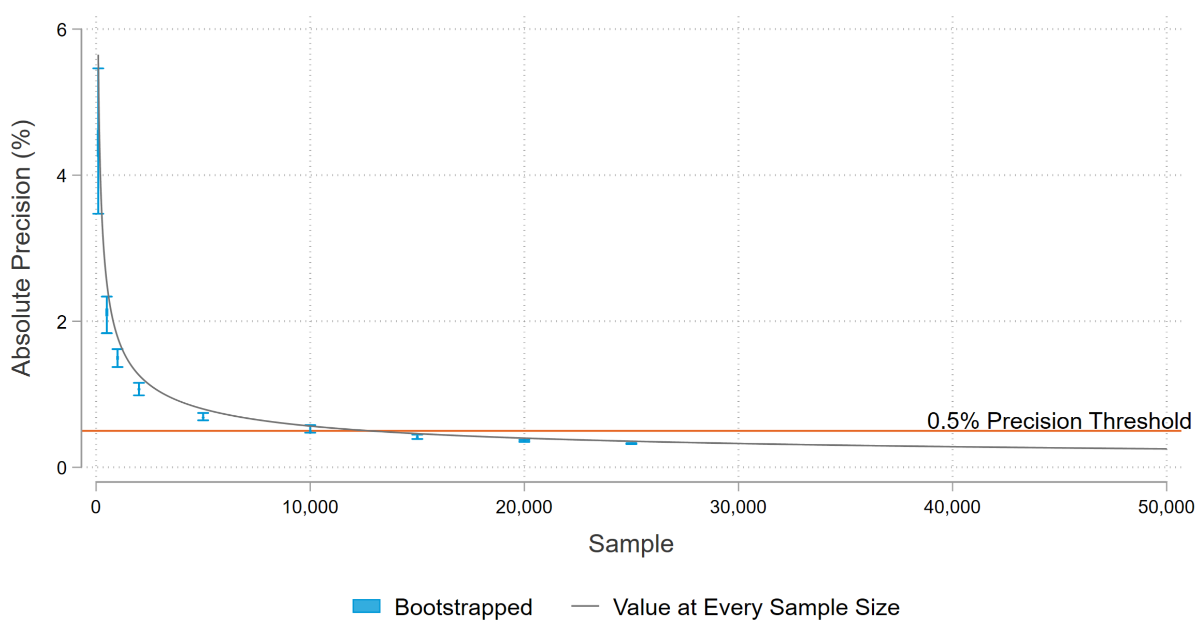
Extremely large control group sizes would be necessary to achieve ±15% relative precision at the 85% confidence level. For BER programs where customer size and consumption patterns are highly variable and expected percent impacts are smaller, 85/15 is likely impossible.

The ± 0.5% absolute precision requirement is for program design and not necessarily ex post savings estimates (although differences between the two should be minimal). EDC ECs should include a description of the data and methods utilized and the results of their expected precision calculations in their EM&V plans or a standalone memorandum for SWE review. If calculations are performed in a reasonable manner and the expected precision of the experiment is at least ± 0.5% at the 95% confidence level, the precision requirement is considered satisfied.

There are several ways to look at the expected absolute precision of an RCT at various group sizes and select group sizes that will meet the required precision level. There are statistical formulas that consider the variability of load data and available population size to calculate the expected standard error of the impact estimate.

EDC ECs can also use a simulation approach known as bootstrapping to approximate the expected precision at various group sizes. The bootstrapping approach works best with at least a two-year period of unperturbed load data (no actual treatment effect). Vendors or ECs then draw hundreds of repeated random samples of the group size of interest and estimate the treatment effect. Since there is no actual effect, the distribution of impacts estimates from repeated iterations will center on zero kWh. The parameter of interest is the standard deviation of the hundreds of estimates, which is what the standard error of a regression model is approximating. Figure 8 shows the expected output from group size investigation (either method). As the control group sizes increase, the expected precision improves.

Figure 8: Hypothetical Sample Size Simulation Output



The relationship is non-linear, which creates diminishing returns for control group sizes past a certain point. While the difference between a 5,000-customer control group and a 10,000-customer control group is dramatic, the precision gain from 35,000 to 40,000 customers is almost negligible. For large HER programs with hundreds of thousands of households, it is unnecessary to have the treatment and control groups sized equivalently.

EDC ECs should never draw samples of homes from the treatment and control groups for impact evaluation. To analyze a subset of participants needlessly erodes the precision of the impact estimate because most statistical packages can easily handle the data volume associated with a large behavioral program. Sampling for customer surveys, or even to some extent for demand reduction analysis, is acceptable.

Opt-Outs and Account Closures

Over time, some homes and businesses assigned to behavioral conservation programs will close their accounts with the EDC. The most common reason is because the occupant is moving, but other possibilities exist. This account churn happens at a fairly predictable rate for an EDC service territory and can be forecasted with some degree of certainty. It is also completely external to the program, so there is no reason to suspect that it happens differently in the treatment and control groups if randomization is done properly. EDC evaluators should include all active accounts for a given month in the analysis and all participation counts used to calculate aggregate MWh savings. Once an account closes, there will no longer be consumption records in the billing data set, so the home or business will be removed naturally from the analysis without any special steps required of the EC.

Many behavioral programs allow treatment group homes to opt-out of receiving HER or BER mailings if they choose. Typically, only a small proportion of the treatment group exercises this option. It is important that EDC ECs do not remove opt-outs from the analysis because doing so could compromise the randomization (control group homes do not have the ability to opt out). The DOE's UMP Residential Behavior Protocol[85] states, "To ensure the internal validity of the savings, opt-out subjects should be kept in the analysis sample." The participant group count should also include customers that have opted out.

Eligibility Criteria

It is important that all eligibility filters be applied when selecting the program population. Then the eligible population should be randomly assigned to treatment and control groups. If randomization into treatment and control groups is performed first and then eligibility filters (e.g., usage requirements, housing type, postal hygiene) are applied, the randomization will be compromised (i.e., the treatment and control households could systematically differ). Even with random assignment to treatment and control occurring after the selection of the eligible population, ECs must still verify that the randomization process was successful, as described in Section 6.1.1.3.

[85] http://energy.gov/sites/prod/files/2015/02/f19/UMPChapter17-residential-behavior.pdf (page 30)

Opt-in Web Portals

While participants in behavioral conservation programs such as HER programs are typically selected randomly as part of an RCT, participants in a web portal, or online audit, program are self-selected. In other words, participation is a choice rather than the result of a randomized process. If an opt-in web portal program is not implemented as an RED, ECs should leverage a matched comparison group of non-participants to evaluate energy savings where:

- Web portal participants that are also in an HER treatment group are matched with non-participants in the same HER treatment group. Otherwise, savings related to HER delivery could be double counted in the web portal analysis.
- Web portal participants that are not in an HER treatment group are matched with other residential accounts that are not in an HER treatment group. The matched comparison accounts could be part of an HER control group but could also come from the general residential population.
- Matching dimensions may include annual energy consumption, hot and cold weather sensitivity, house type, house age, geographic location, and any other variables ECs deem useful.

After developing the matched comparison group, ECs should estimate the verified energy savings using either a lagged dependent variable model or a lagged seasonal model. Section 6.1.1.5 discusses these regression model specifications in more detail. The other two regression model specifications introduced in Section 6.1.1.5 (linear fixed effects regression and control group as an explanatory variable) should be avoided for opt-in web portal evaluations due to issues these models encounter when there is variation in treatment timing among the participants. Additional considerations for a web portal program evaluation are provided in Table 30.

Table 30: Considerations for a Web Portal Program Evaluation

Topic	Discussion
Definition of a participant	Any customer who has accessed the portal is considered a participant until their account closes. An account that accesses the portal one time in PY18 should be considered a participant for the entirety of Phase V (and beyond), assuming their account remains open. This approach is akin to the "once randomized, always analyzed" approach for behavioral programs that are delivered as RCTs or REDs.
Definition of the "pre" period	The "pre" period is the 12-month period before a participant accesses the portal for the first time. Whether or not the portal participant receives HERs does not affect the definition of the "pre" period for the web portal analysis.
Precision requirement	EDCs cannot control the rate at which customers opt-in to web portal, so they have less control over the statistical power of the analysis and the precision of the results. Web portals are not subject to the absolute precision requirements for HER or BER program discussed earlier in Section 6.1.1.1. ECs should use the point estimate for gross verified savings regardless of sign or statistical significance.
Outlier detection	Follow the guidance in Section 6.1.1.4.
Estimating annual energy impacts	As noted, ECs should leverage the lagged dependent variable model or the lagged seasonal model. Both are discussed in Section 6.1.1.5.

Topic	Discussion
Estimating peak demand impacts	Follow the guidance in Section 6.1.1.6.
Aggregating impacts	Follow the guidance in Section 6.1.1.7.
Dual participation analysis	Following the guidance in Section 6.1.1.8, ECs should perform a dual participation analysis to prevent double counting savings.
Incremental annual accounting and measure life	Persistence of savings for opt-in web portals has not been studied. ECs should assume a one-year measure life. The decay calculations shown in Section 6.1.1.9 do not apply to opt-in web portals.

In addition to reporting on energy and peak demand impacts, ECs should report on participation (average number of logins per participant, average time between logins, etc.). Although not required, ECs are also encouraged to analyze and report on variation in the treatment effect based on the frequency of visits. In other words, do participants who visit the portal more frequently see greater savings than participants who access the portal less frequently, on average?

6.1.1.2 Cohorts

For mature behavioral programs, it is common for an EDC to add participants to the program at various points in time. This can be done to offset attrition due to natural account churn, to expand the program to additional participants, or to test new treatment strategies. This creates a situation where the behavioral program consists of multiple waves, or cohorts, that were added to the program at different points in time. EDCs should consider each new cohort to be a separate RCT with random assignment of homes to treatment and control. Under no circumstances should participants be added to the treatment group without a corresponding assignment to the control group.

All impact analyses of Act 129 behavioral programs should be conducted at the cohort level. That is, a separate regression model should be specified to compare the usage of treatment and control group homes in the cohort and estimate the treatment effect for that cohort. Once the average savings per home in a cohort are calculated and multiplied by the number of active treatment group homes in the cohort, the resulting MWh impacts can be summed across cohorts to calculate program performance. EDC ECs can perform a weighted average calculation to produce relevant statistics, such as the average annual kWh savings per home or the average percent savings per home, using the number of active treatment group homes as the weighting factor.

Re-Randomization of Existing Treatment and Control Groups

EDCs, their ICSPs, and ECs should consider re-randomizing existing treatment and control groups into new treatment and control groups and assigning new treatment dates if one or more of the following conditions apply:

1. The EDC does not have access to pre-treatment summer and winter Advanced Metering Infrastructure (AMI) data for most homes.
2. The wave was launched in Phase I or Phase II of Act 129.
3. Less than 90% of the treatment and control group homes have a full 12 months of pre-treatment data.

4. The control group has dropped below 5,000 homes.
5. The wave has not been treated for three consecutive program years.

These conditions are meant to serve as triggers for more detailed discussions with the SWE rather than mandatory requirements. The EDCs and SWE should consider IT limitations as well as the potential impact on approved EE&C Plan savings forecasts before finalizing the list of waves to re-randomize and when the re-randomization occurs. Re-randomization could be undertaken at the beginning of a new Act 129 program year or mid-year to allow savings to ramp up in time for the following summer peak demand season. Since this Phase V Evaluation Framework was published in June 2026, it is likely impractical for EDCs and their ICSPs to re-randomize active cohorts for PY18.

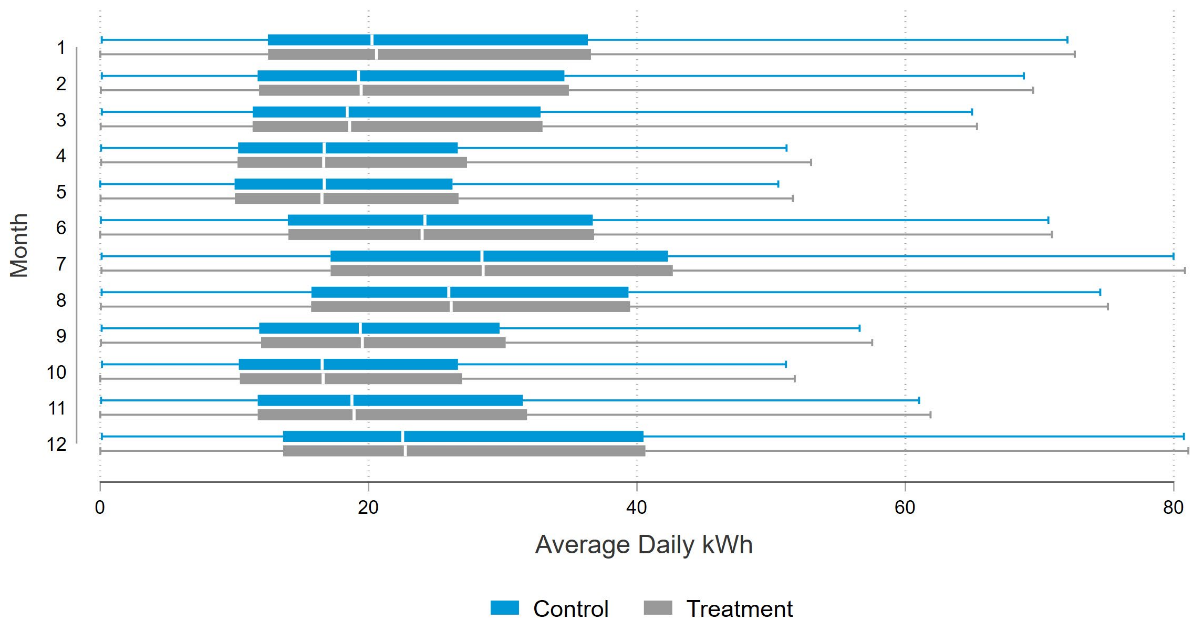
Following re-randomization, ECs should verify that the new treatment and control groups have similar proportions of previously treated households. Re-randomization can be performed in combination with adding more homes or pooling together multiple cohorts that need re-randomization. Re-randomization would effectively restart any treatment these participants receive.

6.1.1.3 Equivalence Testing

Validation of the pre-treatment equivalence of the treatment and control groups is an important feature of behavioral program evaluation because randomization is so critical to the ability to develop unbiased measurements of behavioral program impacts. Randomization can be performed by the EDC, the behavioral program vendor, the EDC EC, or the SWE (if requested). Regardless of who performs the randomization, EDC ECs should carefully examine the equivalence of key characteristics of the treatment and control groups during the pre-treatment period. Electric consumption is the most important characteristic, but if other characteristics (business type, heating fuel, demographics, ZIP code, etc.) are available, they should be examined as well.

The first step of equivalence testing is to perform a visual inspection of the central tendency of the electric consumption of the two groups during the pre-treatment period. Figure 9 shows the results of a successful equivalence check. Notice how monthly consumption varies seasonally but does so in a similar pattern for the treatment and control groups.

Figure 9: Successful HER Equivalence Check



Visual comparisons are an excellent first step and can provide quick indications if the randomization has been compromised. Before considering the treatment and control groups equivalent and the randomization sound, EDC ECs should also perform a statistical test for equivalence. This can be done via a simple t-test or by estimating a random effects model on the pre-treatment period and assessing the significance of the treatment group indicator variable. Another check that should be considered is the relative frequency of estimated meter reads between the control group and the treatment group, performed via a t-test. If these methods indicate a statistically significant difference between the treatment and control groups ($p < 0.10$) and the treatment has not begun, the randomization should be performed again. If the treatment has begun, EDC ECs should alert the SWE immediately to discuss the appropriate corrective action.

When randomization is compromised and the treatment has begun, the SWE will work with the EDC EC to investigate several possible mitigating approaches.

1. Applying filters to the control group that may have been imposed only on the treatment group. For example, perhaps the vendor or mailing house removed all homes with a P.O. Box mailing address from the treatment group, but not the control group. A first step is to apply this filter to the control group and re-examine equivalence.
2. Selecting a matched control group. This technique involves selecting a subset of the cohort control that better resembles the treatment group with respect to observable characteristics (energy use).
3. Using with-replacement propensity score matching (PSM) for the cohort to rebalance the cohort by applying a series of weights for the control customers in that cohort. The EC should conduct the equivalence test again with the weights to ensure the cohort passes.

Those weights are then used in the calculations for consumption savings and peak demand reduction.

4. Provided the parallel trends assumption holds between the treatment and control groups, no adjustments need to be made.

There is a tendency for evaluators to rely too heavily on participant-level fixed effects to control for pre-treatment differences between treatment and control group participants. While a fixed-effects panel regression does help to control for differences in time-invariant characteristics, it is not a panacea for pre-treatment differences in electric consumption. If a fixed-effects panel regression model is estimated for a cohort with statistically significant differences in pre-treatment energy usage that differ over time, the resulting estimate of the treatment effect may be unreliable, and the SWE may object to EDCs claiming savings toward Act 129 compliance reduction goals.

6.1.1.4 Data Management

For EDCs that have advanced metering infrastructure/automatic meter reading (AMI/AMR) in place for all customers and the capability to provide that data to ECs for processing, the data management process for behavioral program analysis is straightforward. Because EDCs have records of the hourly or daily consumption within each home or business, all participants can be easily placed on a uniform basis for analysis. To summarize the March consumption for a given home, the EDC EC simply needs to sum the hourly or daily kWh records from March 1 to March 31. While hourly or daily analysis can yield useful insights (particularly regarding demand reduction, as discussed in Section 6.1.1.6, monthly estimates of the behavioral impacts are sufficiently granular to estimate consumption reductions for Act 129 compliance filings.

For EDCs with traditional mechanical revenue meters, or where AMI/AMR data retrieval would prove burdensome to IT resources, monthly billing data will be starting point for behavioral analysis. With utility billing data, usage is not measured within a standard calendar month interval. Instead, billing cycles may be a function of meter read dates and vary across accounts. Since the interval between meter readings varies by customer and by month, EDC ECs need to calendarize the usage data to reflect each calendar month so that all accounts represent usage on a uniform basis for analysis. The calendarization process includes expanding usage data into daily usage, splitting the bill cycle's usage uniformly among the number of days between meter reads, and assigning them to calendar months. The average daily usage for each calendar month is then calculated based on the days of an individual calendar month.

Occasionally, EDCs will miss a scheduled meter read and estimate the consumption in the home or business during the bill cycle. Once the meter is read again, the customer is billed for the difference between the actual usage for the two-month period and the estimated bill from the first month. EDCs should make sure to delineate actual and estimated reads in the data provided to the EC for analysis. When such data is calendarized for analysis, ECs should sum the consecutive estimated reads together with the first actual read that follows and divide that aggregated use across the number of days since the previous actual read. This will yield the average value in the data calendarization. Table 31 provides an example. For all days between February 16 and May 15, the consumption within the home is assumed to be 38.2 kWh (3,400 kWh ÷ 89 days). Although this approach simplifies consumption patterns considerably, it eliminates the possibility that EDCs' estimated meter reads bias the estimated treatment effect.

Table 31: Estimated Meter Read Calendarization Example

Meter Read Date	Days in Cycle	Estimated or Actual	Billed kWh	Average Daily kWh
2/15/2026	30	Actual	1,500	50
3/15/2026	28	Estimated	1,100	38.2
4/15/2026	31	Estimated	900	
5/15/2026	30	Actual	1,400	

Outlier Detection and Removal

Occasionally EDC billing data will include implausible consumption amounts for homes or businesses that should be removed prior to analysis. Outlier detection should be symmetrical and remove both unrealistically high and low values. Only a small number of data points (less than 1%) should be removed. If more than 1% of the observations in the data set are being flagged for removal, this indicates a utility-side data issue and the SWE should be consulted.

6.1.1.5 Model Specification

There are four general classes of regression model specifications that can be used to estimate the verified energy savings from behavior-based conservation programs. Each model compares the differences in energy consumption between the treatment group and the control group in the treatment period with an adjustment mechanism to account for any observed differences in the pre-treatment period. Although the intent is the same, the models operate in slightly different ways.

1. **Linear Fixed Effects Regression (LFER) Model**. Also referred to as a difference-in-differences (DiD) regression, LFER models estimate the ATE on an absolute basis (kWh). This model has been the most widely used approach to estimate behavioral savings and is the recommended approach in SEEA's protocol for the EM&V of Residential Behavior-Based Energy Efficiency Programs.[86]
2. **Lagged Dependent Variable (LDV) Model**. The LDV model is referred to as a post-only model because only observations from the post-treatment period are included in the regression. Instead of using both pre and post data in the regression, the LDV model uses each customer's energy use in the same month during the pre-treatment period as an explanatory variable. The LDV model estimates the ATE on an absolute basis (kWh).
3. **Lagged Seasonal (LS) Model**. This model is like the LDV but uses pre-treatment consumption data for each home slightly differently. Instead of creating a single lag term, the LS model contains three lag variables: one for average usage (all months), one for

[86] Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations. State and Local Energy Efficiency Action Network. May 2012. Weblink

average summer usage, and one for average winter usage. The LS model estimates the ATE on an absolute basis (kWh).

4. **Control Group as Explanatory Variable (CGEV) Model**. This model identifies the effect of treatment in the post-period by keeping only the experimental group in the dataset and including the average control group consumption as an explanatory variable. The model estimates usage using a fixed effects panel regression with the average daily usage of the control group and a post-period indicator as the explanatory variables. The control group average daily usage variable explains 99% of the variation in the experimental group because the two groups experience the same weather, day of week and other factors. This isolates the impact of treatment in the post period by estimating the effect of the post indicator.

Each of these models has advantages and disadvantages, which are discussed in more detail below. Because of the inherent variability in customer electric consumption, any model will need to isolate the effect (energy savings) from the noise. Because of the different mechanisms by which each model controls for customer characteristics and separates the program effect from the noise in the data, estimating these four models on the exact same behavioral program data set will produce at least slightly different results. To avoid the temptation of estimating multiple models and selecting the approach with the most favorable savings estimate, EDC ECs must specify the model specification that will be utilized to calculate savings in their EM&V plans and provide justifications for their choice.

When multiple models provide similar estimates, the results are considered robust and all stakeholders can be more confident that the estimated savings accurately reflect the true reduction in electric consumption achieved by the program. While EM&V plans need to explicitly state the model specification that will be used to calculate compliance savings, ECs are encouraged to estimate additional models or variants of the same model (e.g., with and without weather terms) to investigate the robustness of the primary model. If the primary model produces inconsistent findings compared to a series of alternative specifications, EDCs may wish to propose to the SWE that a different primary model be used for subsequent program years.

Technical Guidance on Behavioral Models

The basic form of the LFER model is shown in Equation 12. Monthly energy consumption for treatment and control group customers is modeled using an indicator variable for the month of the study, a treatment indicator variable, and account-level fixed effects:

Equation 12: LFER Model Specification

$$kWh_{imy} = \beta_i + \sum_{m=1}^{12} \sum_{y=2011}^{2031} (\beta_{my} * I_{my}) + \sum_{m=1}^{12} \sum_{y=2011}^{2031} (\tau_{my} * I_{my} * treatment_{imy}) + \varepsilon_{imy}$$

Table 32 defines the model terms and coefficients in Equation 12.

Table 32: LFER Model Definition of Terms

Variable	Definition
kWh_{imy}	Customer i's average daily electric usage in month m of year y.
β_i	The intercept term for customer i plus the fixed effect term. Equal to the mean daily energy use for each customer.
I_{my}	An indicator variable that equals one during month m of year y, and zero otherwise. This variable models each month's deviation from average energy.
β_{my}	The coefficient on the month-year indicator variable.
$treatment_{imy}$	The treatment variable. Equal to one when the treatment is in effect for the treatment group. Zero otherwise. Always zero for the control group.
τ_{my}	The estimated treatment effect in kWh per day; the main parameter of interest. Estimated separately for each month and year.
ε_{imy}	The error term.

An advantage of the LFER model is that time-invariant characteristics (both observed and unobserved) are excluded from the model through the household-level fixed effects term. This is desirable if pre-treatment differences in consumption between the treatment and control groups are present. Although the LFER model does not completely correct for randomization issues, it is the most robust choice when the equivalence of the groups is questionable and pre-treatment differences in consumption are observed.

The drawback of the LFER model is that it is less precise because the household-level fixed effects term relies exclusively on within-customer variation. The explanatory powers of time-invariant characteristics (such as demographics) are lost because those terms are eliminated from the model.

Equation 13 shows the basic form of the LDV model. Unlike the LFER model specification, all accounts share a common intercept (βo) in the LDV model. Although a year of pre-treatment data is still necessary, the model is estimated exclusively using post-treatment observations (post-only). The LDV model also uses a different approach to address the uniqueness of customers. The average daily energy consumption from the month of interest prior to treatment ($kWh_{i,m,y-n}$) is used as an independent variable. Additional time-invariant explanatory variables can also be included in the LDV model to produce more precise estimates or facilitate segmentation of results by sub-groups of interest.

Equation 13: LDV Model Specification

$$kWh_{imy} = \beta_0 + \sum_{m=1}^{12}\sum_{y=2011}^{2031}\left(\beta_{my} * I_{my} * kWh_{i,m,y-n}\right) + \sum_{m=1}^{12}\sum_{y=2011}^{2031}\left(\tau_{my} * I_{my} * treatment_{imy}\right) + \varepsilon_{imy}$$

Table 33 defines the model terms and coefficients in Equation 14.

Table 33: LDV Model Definition of Terms

Variable	Definition
kWh_{imy}	Customer i's average daily energy usage in bill month m in year y.

β_0	Intercept of the regression equation.
I_{my}	An indicator variable equal to one for each monthly bill month m, year y, and zero otherwise. This variable captures the effect of each billing period's deviation from the average energy use over the entire time series under investigation.
β_{my}	The coefficient on the bill month m, year y indicator variable.
$kWh_{i,m,y-n}$	The billed kWh for customer i in bill month m in the year prior to the assignment to treatment condition. Term n represents the number of years home i has been in the program. This term controls variability in customer characteristics such as home size and heating fuel.
$treatment_{imy}$	The treatment indicator variable. Equal to one when the treatment is in effect for the treatment group. Zero otherwise. Always zero for the control group.
τ_{my}	The estimated treatment effect in kWh per day per customer; this is the main parameter of interest. Estimated separately for each month and year.
ε_{imy}	The error term.

A major advantage of the LDV model is that it is more precise than an LFER model because it can be estimated via ordinary least squares (OLS) regression and can leverage both within-participant and between-participant variation. The drawback of the LDV model is that it is more sensitive to equivalency issues. If properties like weather sensitivity or heating fuel are correlated with the assignment to treatment, omitted variable bias can lead to unreliable estimates using the LDV model. EDC ECs should only use post-only models when the treatment and control groups are balanced on usage and selection criteria.

Equation 14 shows the basic form of the LS model. Unlike the LDV model, the LS model is summed over three seasons: pre-treatment, summer, and winter. The definition of summer and winter periods is flexible and should be defined by the evaluator.

Equation 14: LS Model Specification

$$kWh_{imy} = \beta_0 + \sum_{season}^{pre,sum,win} \sum_{m=1}^{12} \sum_{y=2011}^{2031} \left(\beta_{mys} * I_{my} * kWh_{s,i,y-n}\right) + \sum_{m=1}^{12} \sum_{y=2011}^{2031} \left(\tau_{my} * I_{my} * treatment_{imy}\right) + \varepsilon_{imy}$$

Table 34 defines the model terms and coefficients in Equation 14.

Table 34: LS Model Definition of Terms

Variable	Definition
kWh_{imy}	Customer i's average daily energy usage in bill month m in year y.
β_0	Intercept of the regression equation.
I_{my}	An indicator variable equal to one for each monthly bill month m, year y, and zero otherwise.
β_{mys}	The coefficient on the bill month m, year y indicator variable interacted with season s.
$kWh_{s,i,y-n}$	Average daily usage for customer i in the pre-treatment season. Pre is defined as the full year, while summer includes the average daily usage from June–September and winter includes the average daily usage from December through March

$treatment_{imy}$	The treatment indicator variable. Equal to one when the treatment is in effect for the treatment group. Zero otherwise. Always zero for the control group.
τ_{my}	The estimated treatment effect in kWh per day per customer; the main parameter of interest. Estimated separately for each month and year.
ε_{imy}	The error term.

The LS model shares many of the advantages and disadvantages of the LDV model. It can be estimated via OLS and produces more precise impact estimates than the LFER model and slightly more precise estimates than the LDV model. Like the LDV model, the LS model is poorly equipped for pre-treatment differences between the treatment and control groups. EDC ECs should only use post-only models when equivalence tests indicate that the randomization for a cohort is uncompromised.

Equation 15 provides the model specification for the CGEV model.

Equation 15: CGEV Model Specification

$$kWh_{imy} = \beta_i + \beta_c * Ctrl\_kWh_{my} + \sum_{m=1}^{12} \sum_{y=2011}^{2031} \tau_{my} * I_{my} * post_{imy} + \varepsilon_{imy}$$

Table 35 defines the model terms and coefficients in Equation 15.

Table 35: CGEV Model Definition of Terms

Variable	Definition
kWh_{imy}	Treatment customer i's average daily electric usage in month m of year y.
β_i	The intercept term for customer i plus the fixed effect term. Equal to the mean daily energy use for each customer.
β_c	The coefficient for the control customer's average daily usage.
$Ctrl\_kWh_{my}$	The average control customer's average daily use in month m of year y.
I_{my}	An indicator variable that equals one during month m, year y, and zero otherwise. This variable models each month's deviation from average energy.
β_{my}	The coefficient on the month-year indicator variable.
$post_{imy}$	The post-treatment variable. Equal to one when the treatment is in effect and zero otherwise.
τ_{my}	The estimated treatment effect in kWh per day per customer; the main parameter of interest. Estimated separately for each month and year.
ε_{imy}	The error term.

Like the LFER model, this model includes participant-level fixed effects that eliminate any time-invariant characteristics from the estimation. However, the panel data in this model only includes treatment customers during the pre- and post-treatment periods. Control group usage during these timeframes is included only as an explanatory variable. The intuition for this model is that exogenous changes in usage are accounted for in the correlation between control group usage and treatment group usage, which is established in the pre-treatment period. The underlying assumption is that treatment does not change this relationship, which should be established based on a statistically equivalent control group. This method is particularly useful for quasi-experimental designs (e.g. RED) and is best-suited for instances like the opt-in portal rather than for RCTs.

Table 36 provides a summary of the strengths and weaknesses of the four classes of regression models discussed in this section.

Table 36: Summary of Model Pros and Cons

Model Specification	Advantages	Disadvantages
LFER	Best equipped to net out pre-treatment differences in energy consumption.	Less precise because between-participant variation is not used.
LDV	Estimates are more precise than LFER because both within- and between-participant variation is used. Easy to segment results by subgroups of interest.	Susceptible to omitted variable bias if treatment assignment is correlated with factors that affect energy consumption.
LS	Most precise, on average.	Occasionally produces erratic estimates.
CGEV	Less susceptible to pre-treatment differences in usage.	Less commonly used in industry evaluations to date.

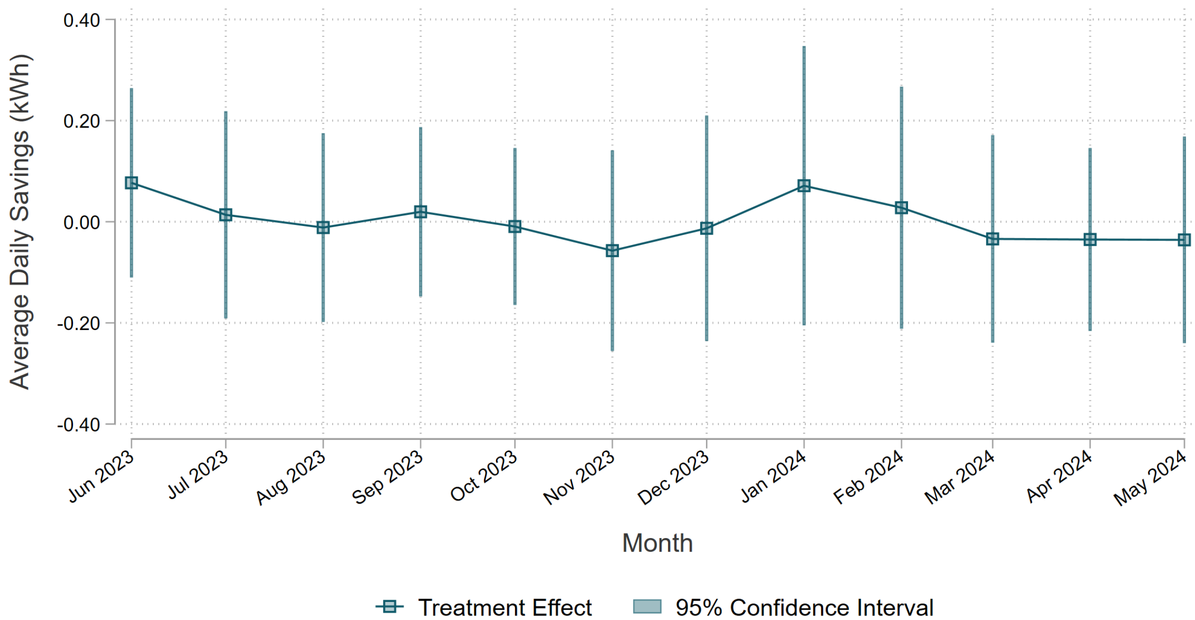
Monthly and Annual Impact Estimates

In each of the model specifications provided in the previous section, the parameter of interest (treatment) is interacted with an indicator variable (month dummies) to produce monthly estimates of the treatment effect (daily kWh savings). EDC ECs should use treatment/time dummy interaction variables to implement this approach when calculating verified savings from behavioral programs. In addition to providing useful information about the saving impacts by time period, monthly (or annual) modeling is important for accurate measurement of program achievements toward compliance goals. When the treatment indicator variable is not interacted with a time-series variable, the result is a cumulative model that estimates the ATE since the inception of treatment for that cohort. This is problematic for Act 129 compliance assessment because many behavioral cohorts have been in place since previous Phases.

Consider an example where a HER cohort began receiving HERs at the beginning of PY14 (June 2022). If, at the end of PY18 (May 2027), an EC estimated a cumulative regression model using a standalone treatment indicator variable, the coefficient would represent the ATE for PY14, PY15, PY16, PY17, and PY18. If the treatment effect grew over time, which many evaluation studies have found, the PY18 savings from the program would be understated.

If ECs prefer, a program year indicator variable can be used in place of the monthly indicator variables. Although the ability to examine seasonal variation in the treatment effect would be lost, the impact estimate would be specific to the Act 129 program year being evaluated. EDC final annual reports should use graphics or tables like Figure 10 to summarize the performance of the behavioral offering over the Program Year. Presenting the confidence interval associated with impacts is encouraged and should be based on clustered robust standard error.

Figure 10: Monthly Impact Estimates with Uncertainty



EDCs should also consider presenting behavioral savings on a percentage basis. Percent impacts can be calculated using Equation 16 and can help normalize impacts to account for the fact that homes and business use different amounts of energy by month, and periods with the highest absolute (kWh) savings may or may not show the greatest savings on a relative basis.

Equation 16: Percent Savings Calculation

$$\%\ Savings = \frac{Average\ kWh\ Savings\ per\ Home}{Average\ kWh\ Usage\ of\ Treatment\ Group + Average\ kWh\ Savings\ per\ Home}$$

Finally, an annual savings measurement is needed for program lifetime and incremental first-year savings calculations. Annual savings are simply the sum of savings from each month that the program was active in the program year. The equations for lifetime savings and incremental annual savings are detailed Section 6.1.1.9, and rely on the annual savings for a given program year for the average account in the relevant cohort or program.

Inclusion of Weather

The model specifications presented earlier in this section do not include weather variables such as temperature, heating degree days, cooling degree days, humidity, etc. One useful feature of the RCT design, if implemented correctly, is that the control group faces weather conditions identical to those of the treatment group, so it is not necessary to include weather variables in the model specification. While not necessary, weather variables can have significant explanatory power for electric consumption and including them in the model may improve precision. EDC ECs are free to include or exclude weather variables from the model specification. This decision should be made in advance and documented in the EM&V plan submitted to the SWE.

6.1.1.6 Peak Demand Impacts

Each of the EDCs has a Phase V PDR target that may be met with coincident demand reductions from energy efficiency or daily load shifting. While EDCs have always been required to produce estimates of the PDRs associated with their HER programs, additional rigor is expected for Phase V because of the compliance target. The Pennsylvania TRM includes definitions for both summer and winter peak demand savings. The TRM defines summer peak demand impacts as the average reduction in electric consumption from 2:00 p.m. to 6:00 p.m. Eastern Daylight Time on non-holiday weekdays during June, July, and August. The TRM defines winter peak demand impacts as the average reduction in electric consumption from 7:00 a.m. to 9:00 a.m. and 6:00 p.m. to 8:00 p.m. Eastern Daylight Time on non-holiday weekdays during January and February. Although behavioral demand impacts are generally small on a per-home or per-business level, when aggregated across thousands of participants, the reductions become meaningful. When selecting an impact approach for peak demand impacts, EDCs and their evaluators should seek to balance level of effort (and cost to rate payers) with the value provided by accurate demand impact estimates based on the specifics of metering infrastructure, IT capabilities and staff bandwidth, and expected savings magnitude.

Preferred Methods for Calculating Peak Impacts

All four EDCs have hourly or sub-hourly revenue meters for most program participants and the IT capabilities to retrieve the data for analysis. This allows evaluators to perform an actual ex post analysis of demand impacts by comparing treatment and control group loads. The models described above can, with a few adjustments, be used to estimate demand impacts. Average hourly demand (kW) becomes the dependent variable instead of average daily kWh.

While all EDCs have had the necessary metering infrastructure for several years, data volume can still be a constraint for EDC staff tasked with pulling interval data or ECs tasked with processing the data for analysis. To the extent possible, the SWE Team recommends estimating peak period impacts at least once in Phase V evaluations, preferably early in the cycle, and referring back to these values for later evaluations. Winter peak demand was not estimated in Phase IV, so evaluators will need to perform a winter peak demand analysis in PY18 or report unverified savings until the analysis can be completed.

Several methods exist to estimate peak period impacts, dependent upon the cohort characteristics, data extract capability and processing power available. The distribution of behavioral savings across hours of the year is not expected to change dramatically from year to year as the allocation will generally be a function of the end-uses where behavior is modified and the load shapes of those end uses. One option EDCs may elect to use is to conduct a full AMI analysis (all months and hours) during a program year early in the phase to develop an 8,760 load shape for HER or BER program savings. In subsequent years EDCs could then just apply this load shape to the verified kWh savings for the program year to estimate peak demand impacts and time-differentiated energy.

In all cases when dealing with large-scale evaluation using granular meter data, data management may become a challenge. EDCs and evaluators are encouraged to filter AMI data requests to what is required to estimate summer and winter peak period estimates:

- Limit the data set to January, February, June, July, and August in the pre- and post-treatment periods
- Exclude Saturdays, Sundays, and holidays
- Select records only from hours ending 8 through 9, and 15 through 20

The peak period estimates from this filtered dataset can be used to construct a ratio of annual energy savings to peak demand impacts for use in subsequent years. This energy to demand factor approach relies on significantly less data manipulation and transfer but does not produce an 8,760 load shape of HER- or BER-related savings.

If data management still proves burdensome to EDC staff and ECs, it is possible to perform the peak demand impact analysis on a sample of participants from the treatment and control groups. If this situation arises, EDC ECs should notify the SWE to determine an acceptable degree of sampling based on the limitations in place. While EDCs are not precluded from estimating program impacts using AMI data for all program years, it may be more practical to rely on billing data analysis and the 8,760 load shape (or the ratio of annual energy savings to peak demand impacts) for subsequent years.

6.1.1.7 Aggregate Impacts

Calculation of aggregate MWh or summer and winter MW impacts from behavioral programs is conceptually straightforward and shown in Equation 17. Starting with the ATE τ (measured in kWh/day and estimated separately by month), EDC ECs simply multiply the ATE by the number of days in each month and the number of active homes in the treatment group during the month.

Equation 17: Aggregate Impact Estimates

$$MWh\ Saved_{PY18} = \sum_{m=1}^{12} \tau_{my} * Days_{my} * Tx\ Accounts_{my}$$

Aggregate impacts should be calculated separately for each cohort and season in a behavioral program and then summed to arrive at an estimate of program performance by season. Treatment group homes that opt out should not be excluded from the impact estimation or participation counts. "Once randomized, always analyzed" is a useful motto for behavioral analysis. Counts should be based on the number of treatment group accounts that have consumption data for the month of interest. Accounts that have closed or moved will not have billed usage and will naturally remove themselves from both the estimation and the count of active participants.

6.1.1.8 Dual Participation Analysis

Exposure to behavioral program messaging often motivates participants to take advantage of other EDC EE&C programs. In fact, many EDCs will include promotional material on other programs within an

HER or BER. This creates a situation where the treatment group participates in other EE&C programs at a higher rate than control group homes. The UMP on residential behavior evaluation states,[87]

> *When a household participates in an efficiency program because of this encouragement, the utility might count their savings twice: once in the regression-based estimate of BB program savings and again in the estimate of savings for the rebate program. To avoid double counting savings, evaluators must estimate savings from program uplift and subtract them from the efficiency program portfolio savings.*

The mechanics of the dual participation analysis are somewhat different for upstream and downstream programs.

Downstream Programs

For downstream programs where participation is tracked at the account level, the dual participation analysis can be completed using the following steps:

1. Match the program tracking data to the treatment and control homes by a unique identifier.
2. Assign each transaction to a month based on the participation date field in the tracking data.
3. Exclude any installations that occurred prior to the home being assigned to the treatment or control group.
4. Calculate the daily kWh savings of each efficient measure. This value is equal to the reported kWh savings of the measure divided by 365.25.[88] ECs can choose to apply the realization rate and NTGR for the relevant program year if those values are available at the time of the analysis.
5. Sum the daily kWh impact, by account, for all measures installed prior to a given month.
6. Calculate the average kWh savings per account per day for the treatment and control groups by month. Multiply by the number of days in the month.
7. Calculate the incremental daily kWh from energy efficiency (treatment – control). This value should be subtracted from the treatment effect determined via regression analysis prior to calculating gross verified savings for behavioral programs.

Figure 11 shows the results of a hypothetical dual participation analysis. Both the treatment and control groups gradually accrue additional efficient installations, so the average savings go up gradually over time for both groups. However, the treatment group participates at a higher rate, or completes larger projects on average, so we gradually begin to observe separation in the average kWh savings per home.

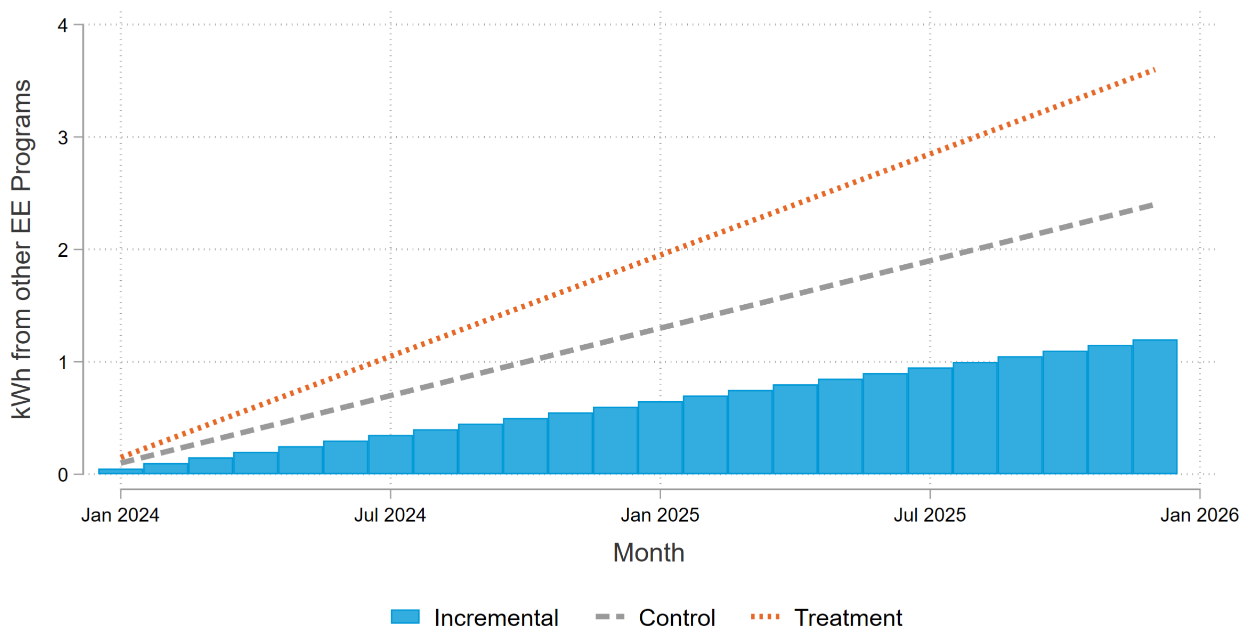
---

[87] Chapter 17: Residential Behavioral Evaluation Protocol. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. Weblink (p. 31).

[88] In practice, most energy efficiency measures save energy at different levels throughout the year based on weather or other factors. The assumption of a flat load shape is intended to simplify the calculations.

This difference, or incremental kWh, is what must be deducted from the behavioral programs' impacts to avoid double-counting.

Figure 11: Dual Participation Analysis Output



Dual participation analysis should be performed and reported separately by cohort in the EDC Final Annual Reports. A long history of tracking data will be needed for cohorts that have been receiving treatment since Phase III of Act 129. If an HER cohort began treatment in January 2017, EDC ECs would need program tracking data and evaluation results for all residential programs back to PY8 to perform the dual participation analysis.

The calculations described above assume that each installed measure will last throughout the period of analysis for the behavioral program. During Phase V of Act 129, long running HER cohorts will see dual participation savings from earlier phases that reach the end of their useful lives. Consider a measure with an EUL of five years installed in 2022. By 2027, the installed appliance has reached the end of its mechanical life and is no longer producing energy savings. EDC ECs are encouraged to account for this phenomenon and remove measures from the dual participation analysis during the months after the end of their useful life.

Winter kW savings were not calculated or reported in Phase III or Phase IV of Act 129, so dual participation calculations for winter peak demand can begin in PY18 even for mature cohorts.

Upstream Programs

Upstream programs present a unique challenge for dual participation analysis because participation is not tracked at the customer level and therefore cannot be tied back to treatment and control group homes for comparison. While incremental uptake of upstream measures by the treatment group has been observed in a number of studies, the size of the effects that are typically subtracted are disproportionate to the evaluation resources required to estimate it.

The UMP for behavioral evaluation recommends that evaluators perform surveys to estimate incremental uptake of upstream measures but acknowledges that "because the individual difference in the number of upstream measure purchases between treatment and control group subjects may be small, a large number of subjects must be surveyed to detect the BB program effect." Given the decline of residential upstream LED lighting programs during the second half of Phase IV, EDC ECs are discouraged from performing surveys to estimate dual participation savings from upstream programs.

While primary research into upstream dual participation is not a good use of evaluation budgets, it is important to correct for the underlying issue. Table 37 provides default values that can be used to calculate a dual participation adjustment factor for upstream lighting offerings. With no mass market upstream lighting program available to the treatment and control group customers in Phase V, the upstream lighting adjustment should reflect historical access to retail lighting programs. To account for the growing separation between the treatment and control groups over time, Table 37 relies on a conditional lookup based on the number of years that a given behavioral cohort had access to an upstream lighting program. Once an EDC's upstream lighting program ends, any cohorts introduced afterward should be assigned an upstream value of 0, while cohorts introduced earlier will have their upstream adjustment factor fixed based on the number of years they had access before the program ended.

Table 37: Default Upstream Adjustment Factors[89]

Years that Cohort Had Access to Upstream Lighting Program	Default Upstream Reduction Factor
1	0.75%
2	1.5%
3	2.25%
4 and beyond	3.0%

The adjustment factors in Table 37 should be applied after the dual participation adjustment for downstream programs is made. The factor can be applied on a monthly or annual basis at the EC's discretion. The following example shows a sample calculation for an HER program cohort in its third year.

$$PY18\ Average\ Impact\ per\ Home = 220\ kWh$$

$$Downstream\ Adjustment = 220 - 4 = 216\ kWh$$

[89] Default values were developed via a review of two studies that used primary data collection with large sample sizes to estimate a dual participation adjustment for upstream lighting. A 2012 PG&E evaluation found values larger than those in this table. Weblink

$$Upstream\ Adjustment = 216 * (1 - 0.0225) = 211.14\ kWh$$

Act 129 evaluations of residential upstream lighting programs have consistently found cross-sector sales of products to non-residential customers. Based on these findings, EDC ECs should apply the adjustment factors shown in Table 37 to BER program results unless surveys or other primary research is conducted to estimate a program-specific dual participation adjustment for upstream programs.

6.1.1.9 Incremental Annual Accounting and Measure Life

Behavioral conservation programs are fundamentally different from a high efficiency piece of equipment that is installed once and then generates savings consistently until it reaches the end of its mechanical life and generates zero savings. One difference is the definition of installation. HER and BER programs rely on repeated messaging to the same homes or businesses to stimulate savings. This creates challenges for applying EUL assumptions and calculating cost-effectiveness.

Phase V energy and peak demand savings goals are based on incremental annual accounting of performance. Each program year, the new first-year savings achieved by an EE&C program are added to an EDC's progress toward compliance. Phase V of Act 129 relies on updated accounting for incremental annual savings and lifetime savings for residential HER programs. As specified in the 2026 TRM,[90] incremental savings for HER programs rely on assumptions about the how program impacts would persist if the treatment were discontinued. These persistence assumptions were developed from a study of residential HER program impact decay at several EDCs,[91] where on average, 31.3% of a program's impact decays each year following the discontinuation of treatment. Persistence begins two years after a cohort is treated. No such study has been done for commercial BERs or opt-in portals; in the absence of a robust study of BER or opt-in portal persistence, evaluators and EDCs should assume a one-year measure life.

Beginning with the 2021 TRM, the PUC sought to better reflect the savings associated with HER programs by more accurately quantifying first year and lifetime savings for residential behavioral programs. A full discussion of the approach for calculating these values can be found in the TRM, however the relevant calculation steps are highlighted below. Note that this accounting mechanism will require EDC ECs to keep careful track of and document program savings and customer counts from prior program years. This calculation does not apply for the opt-in web portals.

Residential Incremental First-Year Savings

Incremental first-year savings from HER programs are defined case wise, as follows.

[90] The 2026 Technical Reference Manual with amendments, at Docket No.M-2023-3044491. Adopted at the September 12, 2024, Public Meeting

[91] Addendum to Act 129 Home Energy Report Persistence Study. Weblink

Equation 18: Incremental First-Year Savings for HER Programs

$$\Delta kWh_Y = ATE_y * Treatment\ Accounts_y * Days_y$$

$$FYSATE_y = ATE_y$$

Where y = 1 or 2, and

$$FYSATE_y = ATE_y - \sum_{x=1}^{x=y-2} FYSATE_{y-x} - FYSATE_{y-x} * Decay * (X - 0.5)$$

$$\Delta kWh_Y = FYSATE_y * Treatment\ Accounts_y * Days_y$$

Where y is greater than 2 and less than 6, and

$$FYSATE_y = ATE_y - \sum_{x=1}^{x=3} FYSATE_{y-x} - FYSATE_{y-x} * Decay * (X - 0.5)$$

$$\Delta kWh_Y = FYSATE_y * Treatment\ Accounts_y * Days_y$$

When y is 6 or more.

In Equation 18, ATE_y is the average daily savings as estimated through the regression analysis described in Section 6.1.1.5 and net of any uplift as calculated in Section 6.1.1.8. Y is the year of the program being evaluated; equivalently, the number of years the program has been in effect for that cohort. The default decay rate is 31.3%.

If the ATE for a given year is negative, two adjustments should be made. First, persistence from this year should be counted as zero. Second, the First-Year Saving Average Treatment Effect ($FYSATE$) should be the maximum value between the ATE and calculated $FYSATE$ inclusive of persistence. No change in $FYSATE$ should be made if the ATE is positive for a given year, even if persistence offsets the ATE entirely.

Residential Lifetime Savings

Lifetime savings for an HER cohort are similarly defined in a case-wise manner. For the first year of the program, lifetime savings are simply the total aggregate program savings. For all future years, lifetime savings takes into account the decay of program impacts over time.

Equation 19: Lifetime Savings for HER Programs

$$\Delta kWh_{Y,lifetime} = ATE_Y * Treatment\ Accounts_Y * Days_Y$$

Where y = 1, and

$$\Delta kWh_{Y,lifetime} = \Delta kWh_Y + \sum_{X=1}^{X=3} \left(\left(FYSATE_Y - FYSATE_Y * Decay * (X - 0.5) \right) * (1 - Churn)^X \right) * Days_{Y+X} * Treatment\ Accounts_Y$$

Where *y* is 2 or more.

The parameters in Equation 19 are defined as in Equation 18. Lifetime savings also account for changes in customer churn, which reflects the change in customer counts in a cohort due to account closures and move-outs. EDCs and evaluators can rely on the default value of 6% for customer churn or can substitute a value specific to the cohort being analyzed. A series of example calculations assuming a 31.3% decay rate, 6% churn rate, 365.25 days per year and an initial treatment size of 50,000 accounts is shown in Table 38.

Table 38: Incremental Annual and Lifetime Savings Example

Year	ATE From Billing Analysis (kWh/HH-day)	Measured Savings Total at Meter (kWh/year)	FYSATE First Year Incremental kWh/HH-day	ΔkWh_y Incremental Annual Compliance Savings (kWh/year)	ΔkWh_y_Lifetime Lifetime Savings (kWh)
Y1	0.050	858,338	0.050	913,125	913,125
Y2	0.055	887,521	0.055	1,004,438	2,453,129
Y3	0.060	910,112	0.014	248,507	606,927
Y4	0.065	926,798	0.024	444,593	1,085,825
Y5	0.070	938,205	0.030	553,063	1,350,741
Y6	0.075	944,906	0.034	613,272	1,497,790
Y7	0.080	947,426	0.030	553,606	1,352,068
Y8	0.085	946,241	0.035	639,714	1,562,367

6.1.2 Process Evaluation

Process evaluations support continuous program improvement and are typically designed to identify opportunities for improvement and successes that can be built upon. Behavioral program delivery is essentially one big data exchange process – from EDCs to vendors, and from vendors to participants. In-depth interviews with key EDC and vendor staff to assess the efficacy of program processes are a recommended activity. Participant surveys can also yield useful insights about the effect of behavioral program messaging on customer attitudes, awareness, recall, and adoption of specific energy-saving behaviors (including some that are identified on HERs and some that are not); and engagement with the reports. Surveys are most meaningful when conducted with randomly selected households or businesses from both the treatment and control groups because the control group responses provide a baseline against which to assess the response patterns of the treatment group. The SWE recommends that EDCs conduct participant surveys with randomly selected households from both treatment and control groups within each participant cohort, then aggregate results to the program level via a weighted average.

6.2 Daily Load Shifting Programs

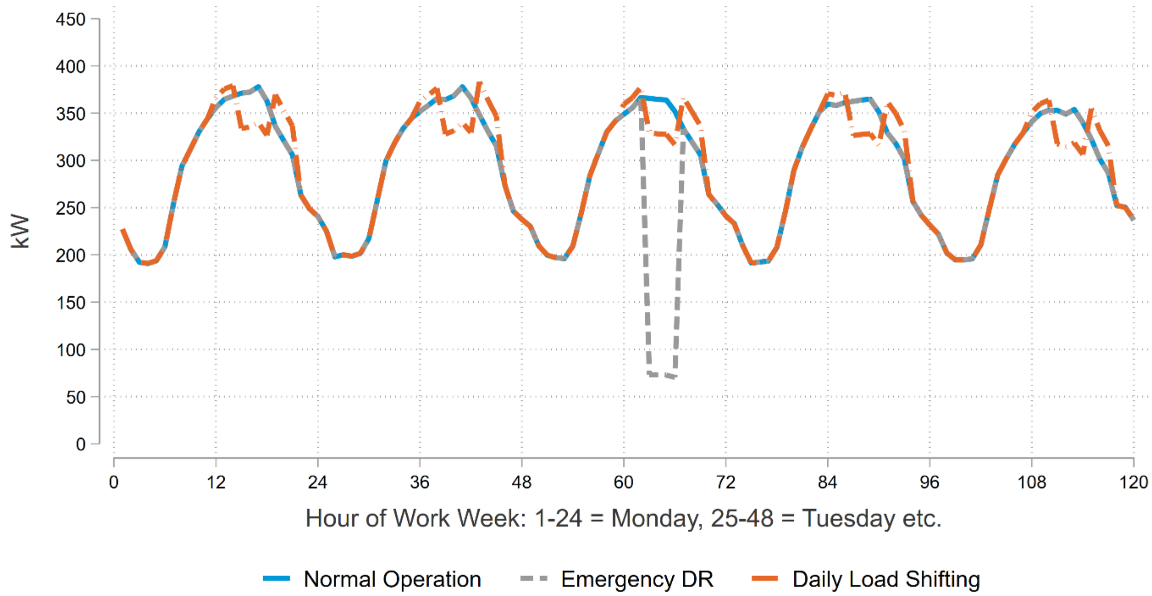
6.2.1 Introduction

This protocol provides guidance to the EDCs and their ECs on estimating summer and winter peak demand impacts for Act 129 daily load shifting programs. The Phase V Final Implementation Order[92] set MWh and MW goals for the four EDCs subject to Act 129 to be achieved by May 31, 2031. A notable feature of the Phase V Final Implementation Order is that it allows the EDCs to achieve peak demand reductions through a combination of energy efficiency, distributed generation, and demand response (DR), leaving the mix up to the discretion of each EDC in its EE&C Plan development process. For DR, Phase V moves away from the traditional event-based framework of Phase III and establishes a daily load shifting framework. Daily load shifting programs work differently than traditional emergency DR programs that are designed to enable large amounts of load shed for a limited number of hours each year. Instead, daily load shifting programs are designed to enable more modest load reductions during the peak period every day, which equates to several hundred performance hours annually.

Figure 12 contrasts the event-based and daily load shifting frameworks. It shows a hypothetical "normal operation" load profile for a commercial business in blue, with an emergency DR load profile in gray and a daily load shifting load profile in orange. While the load reductions are smaller for daily load shifting, they occur in all five weekdays instead of once in this example.

[92] *See Energy Efficiency and Conservation Program* Implementation Order, at Docket No. M-2025-3052826, entered June 18, 2025. Weblink

Figure 12: Daily Load Shifting Versus Event-Based DR



An important consideration for daily load shifting programs is that the treatment is "always on" after a participant enrolls. Although there are not "event days" like with traditional DR programs, every weekday during January, February, June, July, and August is effectively an event day, except for state and federal holidays. New Year's Day, Martin Luther King Jr. Day, and Washington's Birthday (President's Day) would not be a part of the measurement for winter DLS impacts. Juneteenth and Independence Day would not be a part of the measurement for summer DLS impacts. If New Year's Day or Independence Day fall on a weekend, the "observed" date would be excluded from DLS measurement. For example, Independence Day is observed on July 3, 2026, because July 4, 2026, is a Saturday.

After enrollment, there are no non-event, non-holiday, weekdays during the summer or winter season to use for comparison to measure impacts. Even if load shed only occurs on certain weekdays (e.g., only Tuesdays and Thursdays) instead of others for a given participant or program, the remaining weekdays cannot be considered a baseline. In this example, Monday, Wednesday, and Friday would not be eligible baseline days because the site could simply move energy-intensive operations from Tuesday/Thursday to Wednesday/Friday. If those days were used as baseline days rather than performance days, the intraday shifting would overstate the amount of load shift delivered. This feature of daily load shifting programs limits the usefulness of the DR measure characterizations in the 2026 Technical Reference Manual, as the methods in measures 2.9.1 (Direct Load Control [DLC] and Behavior-Based Demand Response Programs) and 3.12.1 (Load Curtailment for Commercial and Industrial Programs) leverage surrounding non-event days to estimate the counterfactual, or reference load. The excerpt below[93]

---

[93] *See Phase V Final Implementation Order* at 157.

from the Phase V Final Implementation Order acknowledges this limitation and cites the need for this daily load shifting protocol.

> *The Commission concurs with PECO that the demand response measures in the 2026 TRM do not provide sufficient guidance to EDCs, CSPs, and evaluation contractors regarding the measurement of daily load shifting program impacts. If one or more EDC elects to propose a daily load-shifting program in its Phase V EE&C plan, we will direct the SWE to update the Pennsylvania Evaluation Framework with specific M&V guidance for the type(s) of programs proposed. EDCs are also encouraged to prepare Interim Measure Protocols for review and approval by the SWE at any time to promote alignment on estimation techniques. The Commission agrees that EDCs should be held harmless for load reductions they sacrifice in service of enhanced measurement accuracy, but this policy should not be used as a back door to deliver event-based DR programs. We will request the SWE author clear guidelines on this issue as part of updates to the Evaluation Framework.*

Importantly, this Implementation Order disposition allows EDCs to withhold a subset of participants or a subset of days from daily load shifting dispatch for measurement purposes. The purpose of "control days" is to allow evaluators to observe participant loads absent the program intervention. For daily load shifting strategies that can reasonably be delayed or "turned off" for a subset of days or participants, the observed loads with the intervention removed can be used to estimate the reference load on days when the intervention is in place, or for participants who received the treatment on the same day. The Implementation Order excerpt above also mentions holding EDCs harmless for control days, so it is important to define what harmless means for EDC verified savings analysis and reporting. Any time an EDC deliberately withholds treatment from a group of willing potential participants temporarily to facilitate measurement, there should be a mechanism to credit the EDC for the forfeited demand reduction. Consider the following illustrative examples.

1. An EDC has a thermostat optimization program with 50,000 participating devices. Devices deliver 0.1 kW of summer peak load reduction, on average, when optimized. If the EDC and its conservation service provider (CSP) elect to withhold 20% (10,000) of devices each weekday of the performance season to act as controls, the estimated load reduction at the meter would be 40,000 * 0.1 = 4,000 kW. However, the withheld devices could have delivered load relief had they not been assigned to a non-optimized state for the day for measurement purposes. When claiming gross verified savings, the EDC would divide the observed load impact by the share of participants that were not withheld and claim 4,000 / (1–0.2) = 5,000 kW.
2. An EDC has a smart water heating program with 10,000 households participating that each have one water heater. On days when curtailment is active, the average winter peak load reduction is 0.2 kW. The EDC implements load shifting on 35 of the 42 non-holiday weekdays in the winter season. The seven control days are used to estimate the counterfactual for the 35 active days. When claiming gross verified savings, the EDC would report the average performance of 2,000 kW measured on the 35 active days with no need to derate compliance savings for the lack of performance on the seven control days. This is mathematically equivalent to assigning 2,000 kW of load reduction to the seven control days as well as the 35 active days and averaging across 42 days.

The availability of control days means that some of the concepts in the 2026 TRM measure characterizations are still applicable to daily load shifting programs. However, implementation of control days is more feasible for some offerings than others. The technology-specific sections of this protocol discuss this issue in more detail. To aid ECs and EDCs in the design of the Phase V EE&C and EM&V plans, the SWE is preemptively providing the M&V guidance for daily load shifting programs called for in Section B.7 of the Phase V Final Implementation Order.

To ensure that the EDCs are not using control days or control groups as a "back door" to deliver event-based DR, the SWE is proposing the following limitations on their use:

1. If an EDC elects to withhold a subset of participants from program dispatch, no more than 25% of participants should be withheld on any given summer or winter weekday.
2. If an EDC elects to withhold a subset of days from program dispatch, no more than 25% of the non-holiday weekdays should be withheld in any given season.
3. If an EDC elects to use both strategies (1) and (2) in the same season, the limits on days withheld and on participants withheld should each be lower than 15%.
4. EDCs and their CSPs may propose alternative limits for withholding participants or days for the first year that a program is offered if they can justify how the alternative limits will improve measurement accuracy.
5. EDCs cannot claim savings from the control group in a Randomized Control Trial with static group assignments. Like the behavioral HER programs discussed in Section 6.1, an EDC incurs no cost for control group members and control groups need not be large enough to materially disadvantage the EDC towards meeting peak demand reduction targets.
6. If an EDC utilizes a recruit and delay strategy where participants are enrolled in the program, but the intervention is withheld for a period of time, the maximum delay period the EDC can claim savings for is one program year (a single summer and a single winter).

Section 3.8 of this evaluation framework states that EDCs and their ECs are not required to conduct ex post evaluations for every EE&C program in all five years of the phase. This guideline also applies to daily load shifting programs and ECs are encouraged to propose a rotating impact evaluation schedule across Phase V. However, the proposed schedule should include a primary impact evaluation for the first year of program implementation and details on how verified gross savings will be claimed for program years that do not receive an impact evaluation. Point #4 above notes that EDCs may propose alternative control day limits for M&V purposes for the first year a program is offered, which may be used to solidify first-year evaluation findings as part of a rotating impact evaluation schedule. The SWE would be supportive of an approach that leans more heavily on control days in the first-year impact evaluation, but fewer control days for impact evaluations later in Phase V and no control days in years when no impact evaluation occurs. For example, an EDC and its EC might propose a 50% withholding strategy for PY18 and a 10% control day assignment for PY19 through PY22, adhering to the 25% limit. The SWE would be supportive of this type of approach as it would allow for greater measurement accuracy in the first-year impact evaluation where less is known about program impacts, and those findings could be leveraged in the following years.

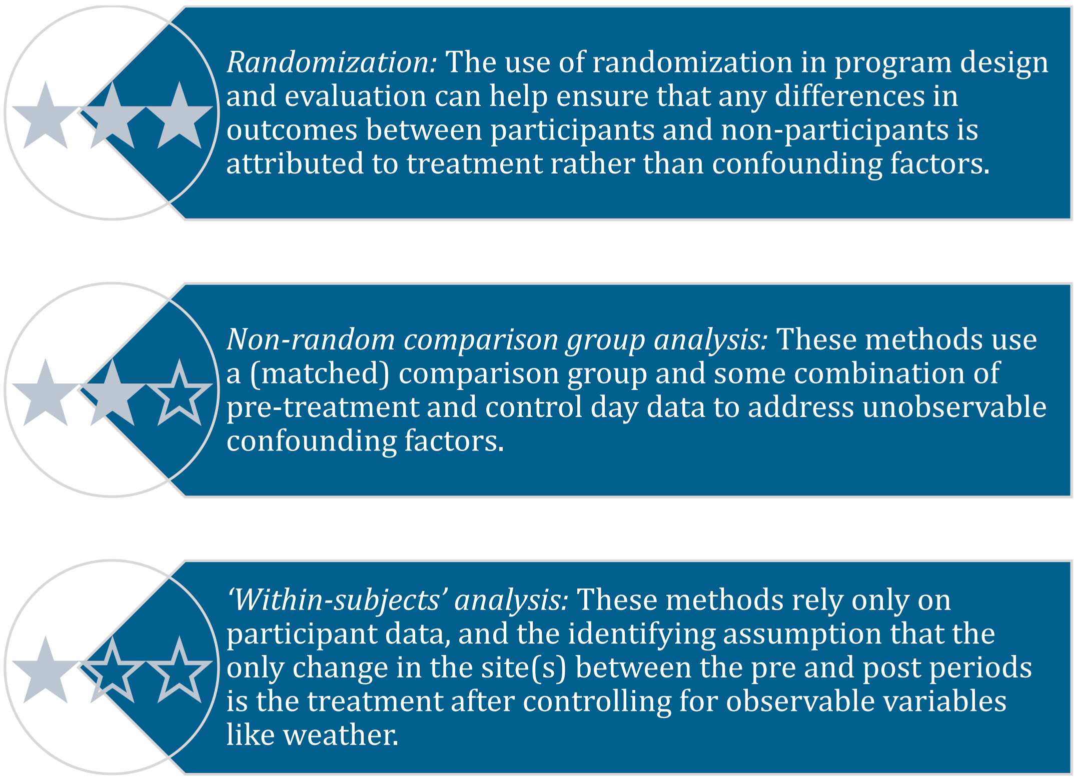
If a DLS program launches at a limited scale during PY18 with plans to expand in future program years, EDCs and their ECs may choose to postpone the first primary impact evaluation until PY19 when the population size is larger. In this scenario, PY18 impacts would be reported as "unverified" until the impact evaluation is completed and the findings could be applied to PY18.

Lastly, the Phase V Final Implementation Order allows EDCs to propose alternative summer and winter performance windows in their EE&C plans, so long as the EDCs provide a rationale for the alternative windows. In addition, the EDCs are free to propose windows that differ based on local distribution needs. Changes in the performance window could have a significant effect on the operations of daily load shifting programs and the applicability of past evaluations. While only PPL proposed an alternative performance in its Phase V EE&C Plan, other EDCs may file an EE&C Plan change and propose one later. The change in performance window, alongside a verification of the anticipated benefits, must be provided as part of a submitted EE&C Plan change.

6.2.2 General Methods

The method by which EDC ECs measure daily load shifting program impacts should be determined according to each program's design and documented in the EM&V plan. The more robust techniques will require close coordination with the ICSP. While the intent of this protocol is to lay out options and provide technical guidance rather than prescribe methods, it is important to establish upfront that all methods are not equally robust. Figure 13 provides a general hierarchy of methods which echoes the methodological preferences expressed by the Commission in the 2026 TRM. The sections that follow discuss each level of the hierarchy in more detail.

Figure 13: General Hierarchy of Methods



6.2.2.1 Randomization

As indicated in Figure 13, the use of randomization is the gold standard for measuring program impacts. For programs like default time-of-use (TOU) pricing, this could take the form of withholding a randomized group of eligible accounts from program enrollment to act as a control group. The EDCs have many years of experience with Randomized Control Trial (RCT) designs from behavioral Home Energy Report programs. With RCT designs, measurement of impacts is accomplished via a straightforward comparison of energy consumption in the treatment and control groups. Withholding a randomized control group is critical for programs like default TOU rates, where enrollment is automatic and the per-participant effect is small.

A Randomized Encouragement Design (RED) is like an RCT except the accounts randomized to the treatment group are offered the treatment rather than assigned to it. RED experiments are often implemented when the randomization of treatment is infeasible. With RED experiments, causal effects can be measured with straightforward techniques such as instrumental variable estimation. Additional details regarding REDs can be found in Chapter 17 of the UMP.[94]

[94] Chapter 17: Residential Behavioral Evaluation Protocol. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. Weblink.

An Alternating Treatment Design (ATD) also leverages randomization, but the treatment status varies across days among enrolled participants rather than across the enrollment process. This powerful "on-off" technique can be applied to opt-in program designs by randomly assigning which participants receive the intervention or act as controls on a given day. The ATD approach works best for mass market offerings where the EDC has direct control over the end use equipment and can automate deployment of the randomized operations plan. Programs that rely on behavioral changes are not well-suited to an ATD because it is difficult to intermittently "turn off" the load shifting practices that participants have been coached to implement. For example, if an EDC program encourages EV owners to delay their charging until after midnight every day, that behavior will likely become ingrained in participants. Alerting a participant that on July 12 it is okay to charge their EV as they would absent the program is problematic because they may not (a) pay attention to the communication, (b) remember what their prior charging behavior was, or (c) bother to override any timers or app settings they have put in place to respond to program messaging. Typically, program offerings with little to no human behavioral components will work the best for an ATD.

6.2.2.2 Non-Random Comparison Group Analysis

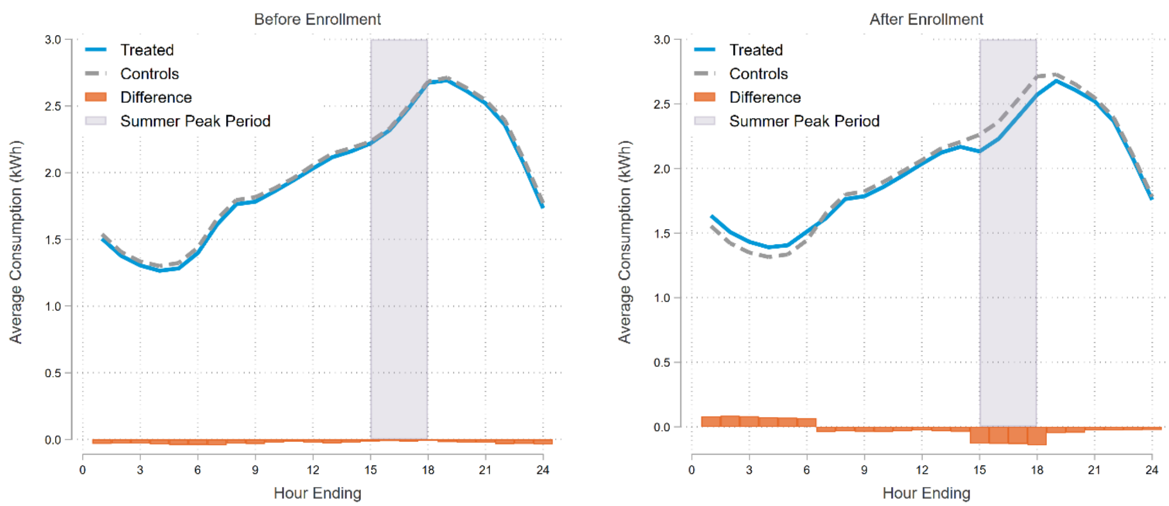
When randomization of the program enrollment mechanism or the assignment of the intervention on performance days is not feasible, a non-random comparison group analysis is often the most robust option. Non-random comparison groups must be carefully constructed to ensure that the treatment group and comparison group are as similar as possible (other than the assignment of treatment). At a high level, constructing a properly matched control group involves: (a) starting with a pool of potential controls that are generally similar to the treated group, (i.e., in the same sector and of a similar size), (b) selecting the accounts from this pool that are the most similar to the treated group based on pre-intervention load patterns and customer characteristics, and (c) verifying the statistical equivalence of the selected group. Under the daily load shifting framework, where the treatment is continuous during the performance season after enrollment, matching should be done with pre-intervention data.

In general, the SWE recommends that the starting pool of potential controls be at least the size of the treated group and ideally many times larger. Starting with a larger number of potential controls can help ensure the treated group is matched with a control group as similar as possible. ECs should request whatever number of similar potential controls and timeframe of data that are necessary for optimal matching. For example, to evaluate a program with 10,000 residential participants first treated in June 2026, an EC might request AMI data for 20,000 residential non-participants from June 2025 through May 2026. The EC should also request AMI data for the program participants for the same period to be used for matching.

Euclidean distance or propensity score "soft" matching on features of the pre-period AMI, such as average hourly loads, maximum demand, or weather sensitivity is often a strong starting point. Leveraging customer characteristics such as location or solar status for "hard" matching may also prove helpful for match quality. Match quality should be assessed on the pre-period only. Statistical equivalence among the treatment and matched control groups can be verified with statistical tests across key dimensions such as average daily consumption and the average demand during the Act 129 peak demand periods. These tests should indicate no statistically significant differences between the groups in aggregate or among sub-groups of interest. Further guidance on constructing matched control groups can be found in Chapter 8 of the UMP.[95]

For programs like opt-in TOU rates, a statistically valid matched control group should be used when estimating impacts. It is important, however, that this control group is constructed based on energy usage patterns before TOU enrollment, as the rate is in place every day after enrollment. See Figure 14 for an example of load patterns for a statistically valid matched control group. In the left panel, which depicts energy usage before TOU enrollment, the average treatment group and control group loads are almost identical, indicating that the matched control group is well-constructed. In the right panel, treated and control loads diverge after enrollment, and this difference is assumed to be the effect of the treatment.

Figure 14: Well-Constructed Matched Control Group Loads



For opt-in TOU rates, evaluation methods using a matched comparison group are always preferred to within-subjects analysis. ECs should not use price elasticities or per-participant impact assumptions from secondary research to estimate impacts unless those assumptions are sourced from a rigorous ex post evaluation of the same program in a prior year of Phase V. If a subset of program participants cannot be included in the analysis for some reason, it is acceptable to estimate impacts for a subset of participants and extrapolate the findings to the population. For example, if an EDC TOU program enrolls homes in the time-varying rate at the time of account creation, there is no pre-treatment data available for matching. Therefore, the TOU effect for these homes would need to be inferred from households where measurement was possible. Similar situations will exist for each of the different DLS

[95] Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. Weblink.

technologies discussed in Section 6.2.3 that require evaluators to "assign" savings to a small subset of participants who do not meet the data sufficiency requirements for inclusion in the analysis.

6.2.2.3 Within-Subjects Analysis

Unless program delivery includes control days where the treatment is withdrawn for a subset of days, a within-subjects analysis is effectively a pre-post model with the identifying assumption that the only change affecting participant load patterns is the program intervention. This pre-post methodology is useful where it is difficult or impossible to withhold treatment for a subset of days or where no suitable matched control can be identified. The use of within-subjects models is well-established in Pennsylvania. Site-specific regression models and IPMVP Option C methods have been used to evaluate energy efficiency impacts for Retrocommissioning, Virtual Commissioning, and custom projects since Phase I. These types of models, when used with hourly or sub-hourly data, are well-suited for evaluating daily load shifting demand impacts as well. ECs are encouraged to leverage any type of data that enhances model accuracy. For Large C&I customers, incorporating weekly or daily indicator variables to capture the periodicity of production schedules may be a useful strategy. For sites where behind-the-meter solar is present, the inclusion of solar irradiance data in the model may help with prediction accuracy. Leveraging non-participant data is possible with the use of granular or class profiles for cases where there is no suitable matched control or randomized control group. An hourly profile by industry type, used on the right-hand-side as a synthetic control[96], often increases prediction accuracy. See Section 6.2.3.4 for more background on the use of granular profiles.

A recruit-and-delay design is a potential enhancement of the within-subjects approach where each participant, or a subset of participants, have their treatment period deliberately withheld for measurement purposes. Delaying optimization until a few weeks or months after a given customer enrolls allows the evaluator access to load data with the treatment "off" for each customer. If enrollment occurs on a rolling basis, delay period data would be available concurrently with other participants who are actively load shifting on the same date. The use of date fixed effects in the model specification would allow evaluators to leverage the observed differences between treated and not-yet-treated participants in the estimation procedure. A recruit-and-delay methodology is useful for offerings with expected behavioral changes from daily load shifting that are difficult to intermittently turn off. Since the assignment of treatment is not random and there is no comparison group of non-participants, the recruit-and-delay design is ultimately a pre-post comparison. However, because the timing and duration of the pre-treatment period is part of the experimental design, it is a methodological enhancement. Since EDCs would forego potential kW savings for the delay period,

[96] The synthetic control method involves adding one or more 8,760 aggregated control group profiles (of which some are publicly available for specific industry types) to the impact model specification as an explanatory variable. This approach relies on these added profiles to construct a synthetic control that leverages the relationship of consumption patterns between the aggregated control group loads and the participant loads during the pre-treatment period to predict participant usage during the post-treatment period. Synthetic controls show comparable results to matched control groups. See Pacific Gas & Electric NMEC Control Group Accuracy Assessment. Weblink.

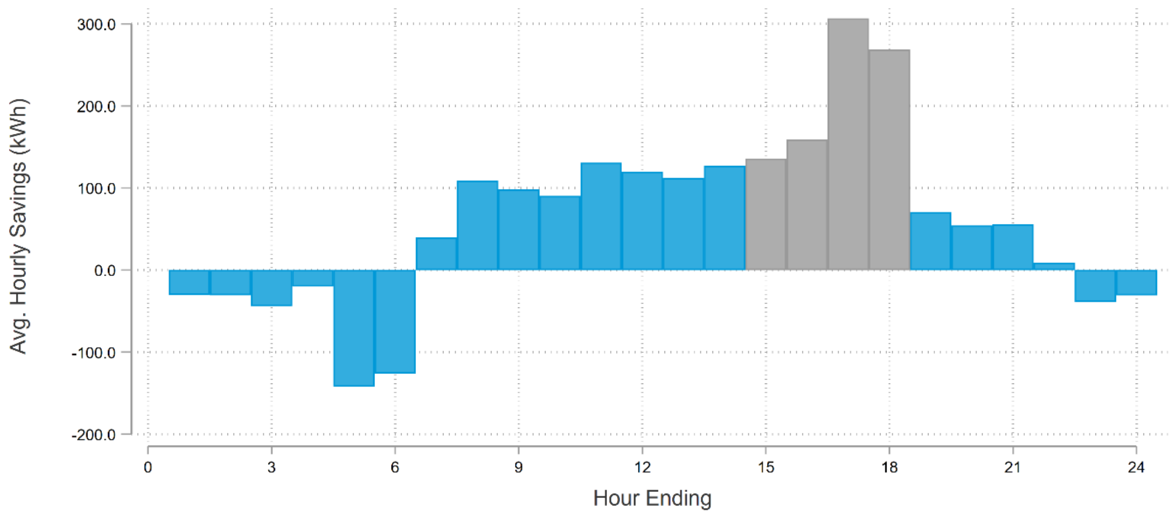
they should be held harmless for the enhancement and credited with the average treatment effect for the duration of the delay period.

6.2.2.4 Energy Impacts

If the reductions in energy consumption during peak periods are not completely offset by energy increases in off-peak periods, a daily load shifting program could have energy impacts in addition to the peak demand impacts it targets. These energy impacts could be positive or negative. For example, battery storage will likely have negative energy impacts because the round-trip efficiency of the battery is less than 100%. While energy impacts are a secondary consideration for daily load shifting programs, they could be non-trivial due to the number of annual performance days each year. EDC ECs should estimate, and report verified gross energy impacts in their annual reports to the PUC. From a methodological standpoint, this means EDC ECs must always include each of the 24 hours of the day in their analysis rather than restricting the analysis to specific hours. ECs should model and report the energy impacts of daily load shifting programs toward Phase V consumption reduction targets regardless of the sign or statistical significance. Positive energy impacts will contribute toward compliance, and negative energy impacts will work against MWh goals. Modeling energy impacts when no effect is expected is often a useful tool for diagnosing model misspecification or other issues. The assessment of energy impacts should be limited to performance days and as such energy impacts that take place outside of performance days (such as Saturday battery charging to recover from Friday load shifting) can be ignored. If a DLS offering is determined to be effectively energy neutral by the evaluator in one program year, they may propose to omit the step and set the energy impacts equal to zero in subsequent years.

Program participants may also elect to implement load shifting strategies alongside energy efficiency measures to maximize bill savings and Act 129 incentives. The SWE anticipates this type of strategy would be most common in a Retrocommissioning or Virtual Commissioning type of program design. Provided that appropriate meter-based methods are used for the energy efficiency analysis, the daily load shifting kW impacts can be bundled and reported alongside the coincident demand reductions of the energy efficiency offering. Figure 15 provides a visual example of a site that, in addition to saving total energy consumption over the course of the day, significantly reduces demand in the peak period by shifting its energy use to earlier in the day (most notably from HE17 and HE18 to HE5 and HE6).

Figure 15: Daily Load Shifting with Energy Efficiency Impacts



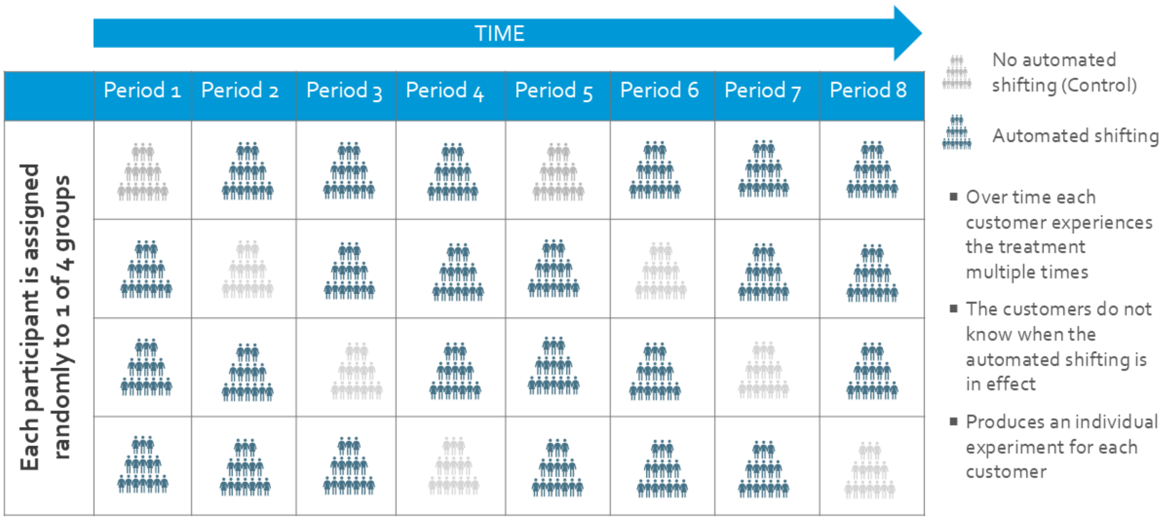
6.2.2.5 Documenting Operations and EM&V Strategies

The EM&V plan should clearly lay out an operations plan that specifies how the day-withholding and/or the participant-withholding strategy will be used to measure program impacts. The operations plan should indicate:

1. The percentage of days that will be withheld from treatment.
2. The percentage of participants that will be withheld from treatment.
3. Documentation on how participants will be grouped for experimentation (i.e., how participants are assigned to groups, the number of participants within each group, and the number of groups)
4. How participants that enroll over the course of the season will be incorporated into the operations plan.

The SWE recommends an accompanying visual that illustrates key facts of the operations plan, such as in Figure 16.

Figure 16: Example Operations Plan With 25% Withholding



The SWE does not envision that sampling should be used extensively for the evaluation of daily load shifting programs. In most cases, analyzing all program participants for all analyses is recommended. However, in instances where the data volume is too large to feasibly include every participant, or in instances where the technology adoption is too close in time to program enrollment to find a suitable baseline period, dropping some participants from the analysis and scaling up impacts may be appropriate.

6.2.3 Technology-Specific Considerations

There are many methods and technologies for enabling daily load shifting for both the residential and non-residential sectors. For the C&I sector, this could include automated building or manufacturing controls that move energy-intensive equipment operation to different hours of the day, then curtail operation during the peak period. For the residential sector, optimization algorithms can be applied to thermostats and water heaters to pre-cool or pre-heat the home or its domestic hot water during the less constrained off-peak hours and curtail cooling load on-peak. Similarly, at-home or fleet EV charging can be optimized such that charging demand is delayed until the less constrained off-peak period, when supply is more abundant, despite what time the vehicle is plugged in. Table 39 details the technologies each EDC plans to offer as part of their daily load shifting program, according to their Phase V EE&C plans. The following sections provide guidance for these technologies. It is not intended to be a comprehensive list and omission from this section should not be interpreted as the SWE or Commission's assessment of suitability of a given offering in Phase V.

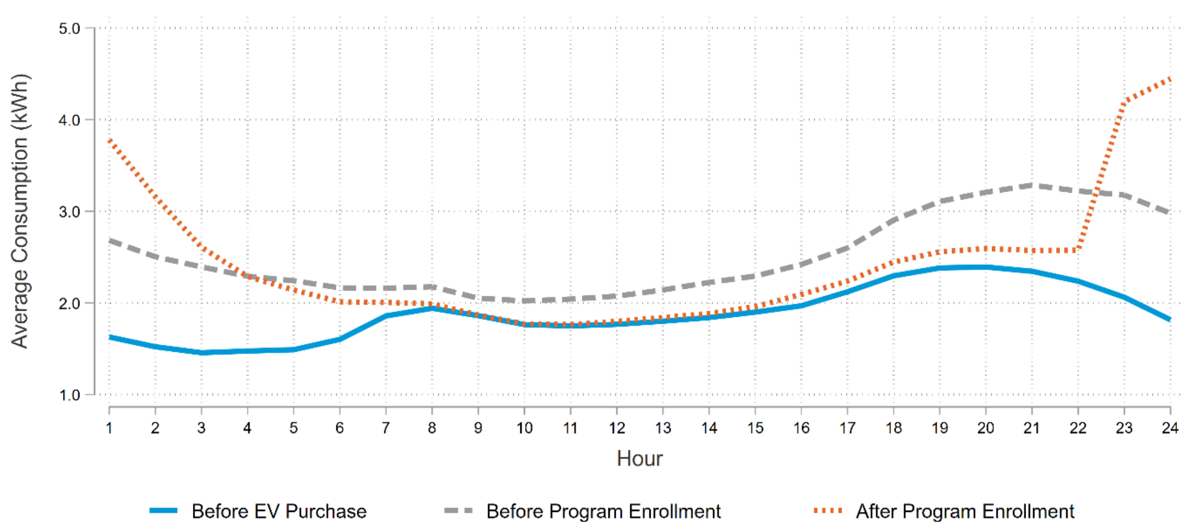
Table 39: Daily Load Shifting Technologies Proposed by EDCs

Technology	Duquesne	FirstEnergy	PPL	PECO
EV Managed Charging	✓	✓	✓	✓
Smart Thermostat	✓	✓	✓	✓
Hybrid Heat Pump Load Shifting Pilot				✓
Battery Storage		✓	✓	✓
C&I Load Shifting	✓	✓	✓	✓
Time-Varying Pricing	✓	✓		✓

6.2.3.1 EV Managed Charging

For EV managed charging or other types of EV load management, EDCs should give special consideration to when program participants purchased their EV. If an EV purchase occurs at the same time as enrollment in the program, this could affect the counterfactual, biasing impacts. Consider Figure 17, which depicts average household load patterns before the adoption of the EV (blue), after EV adoption but before program enrollment (gray), and after program enrollment (orange). To assess the impact of the program directly, the correct counterfactual is household load after EV adoption but before program enrollment (gray dashed line). However, if the program enrollment happens at the same time or very close in time to the adoption of the EV, the only counterfactual available will resemble the blue line, from which an evaluation would conclude inaccurately that the program caused participants to use more energy during the afternoon/evening peak.

Figure 17: Load Patterns Across Pre-Technology, Pre-Enrollment, and Post-Enrollment Periods



The appropriateness of randomized control days for EV managed charging depends on the nature of the intervention and how much behavioral messaging it includes. EV managed charging programs should also be mindful of bill impacts for TOU customers. There could be a break in customer trust if the program did not optimize on certain days and participants charged their vehicle during peak pricing hours. Since managed charging participants are often educated on the best times to charge their EV, they may have these "good habits" ingrained in their charging behavior and might not be able to simply revert to pre-enrollment behaviors just for a control day. In instances like these where pre-enrollment behavior might not be replicable due to the human component of the intervention, a non-randomized comparison group analysis or recruit-and-delay strategy is recommended over an alternating treatment design. Participants should be matched with control customers with an EV but that are not enrolled in the program. In the matching process, special attention should be paid so that EV adoption dates are similar across treated and control. One challenge this approach presents is that while the EV adoption date can be gathered from participants at the time of enrollment, it cannot be gathered directly from non-participants. If the analysis is being done with participant and control AMI data, there are options to address this challenge. One potential solution to this problem that has been used in EV TOU rate impact evaluations in California[97] is to develop an EV detection algorithm to predict the presence of EVs among control group AMI data. The AMI data for sites with EVs often exhibits key unique characteristics that are specific indicators of EV charging, such as high maximum demand on temperate days where air conditioning load is minimal. Once likely EV owners are identified by their unique load profiles, historic AMI data could be analyzed to determine the date on which the site's consumption patterns changed suddenly. For example, if a household has an average maximum demand of 2 kW on mild days in the time period before June 1 and an average maximum demand of 4 kW on mild days in the time period after June 1, an evaluator could reasonably assume the EV charger was installed on June 1.

EV managed charging impacts could also potentially be estimated using data from EV chargers or from vehicle telematics. The challenge with this type of data is that it is typically only available after program enrollment, which makes it challenging to use for estimating the counterfactual charging behavior. Potential remedies include:

- **Creation of a load research sample.** This would entail enrolling a subset of eligible EV owners into a "research only" group where their charging behavior is monitored, but the program does not attempt to shift charging loads away from the peak demand period. The load research group would be used to estimate the baseline for active participants. After a period of 6–12 months, this group could be released to the normal managed charging program experience. The SWE recommends that the EDCs be held harmless for any load research assignments in the same manner as deliberate control days.
- **A recruit-and-delay tactic.** This approach would involve accepting EV owners into the program but not implementing any managed charging features until a baseline charging profile can be collected. The delay period could be as short as a few weeks, but CSP data collection would need to be diligent with respect to the transition date from baseline period to managed charging.

[97] See 2024 Load Impact Evaluation of San Diego Gas and Electric's Electric Vehicles Time-of-Use (TOU) Rates. Weblink. Page 13.

If an EDC plans to use vehicle telematics data to estimate verified gross impacts from its EV managed charging program, the first impact evaluation should include an accuracy assessment where the telematics data is compared to an independent data source such as end-use metering or EV charger data for the same sample of vehicles. Whole-home AMI data cannot be used to assess the accuracy of telematics data as it is impossible to know the true contribution of the EV to the whole-home load. While whole-home AMI could potentially demonstrate that the telematics data is very incorrect (e.g. the telematics-based estimate of charging kWh exceeds household kWh for a non-solar account), it cannot confirm if the telematics data is correct.

6.2.3.2 Storage-Enabled Daily Load Shifting

This section lays out specific considerations for storage-enabled daily load shifting that is (a) behind-the-meter and (b) operated independently of PJM wholesale markets. The most common type of technology used for storage is batteries, either standalone, paired with solar PVs, or potentially part of an electric vehicle in a vehicle-to-grid offering, but similar considerations apply to thermal storage as well. The Commission envisions two kinds of battery programs possibly being implemented in Phase V: (1) subsidy programs where an EDC pays part of the upfront material and installation cost of the battery in order to have control over its charge and discharge, and (2) Bring Your Own Battery (BYOB) programs, where participants who already have storage installed simply enroll in the program and accept EDC incentives to discharge it in a way that is beneficial to the grid. The excerpt below[98] from the Implementation Order proposes an accounting framework for new storage installations and calls on the SWE to develop additional evaluation protocols.

> *Behind-the-meter energy storage measures are a unique offering that received specific discussion in the Tentative Implementation Order. Since EDCs would need to incentivize the upfront capital cost of new battery storage systems or other storage technologies in exchange for an agreement to discharge the resource during the peak demand period, the Commission proposed that EDCs follow an EE accounting framework for new storage installations. Under an EE framework, once the verified summer and winter demand reductions from a storage project are calculated by the EM&V contractor, the reductions can be assumed to persist for the life of the technology and claimed towards goals. Alternatively, if an EDC simply enters into an agreement with a customer with an existing storage system to charge and discharge in a way that contributes peak demand reductions, those impacts would be claimed like other load-shifting programs. The Commission also proposed that the Phase V SWE add a storage protocol to the Pennsylvania Evaluation Framework.*

For new batteries installed under a subsidy program, the "EE framework" described in the Implementation Order means that the discharge of the battery is counted directly as a program impact. This means EDCs should secure access to inverter data as a condition of the program incentive and ECs should use inverter data to measure verified gross peak demand impacts. Battery discharge data does

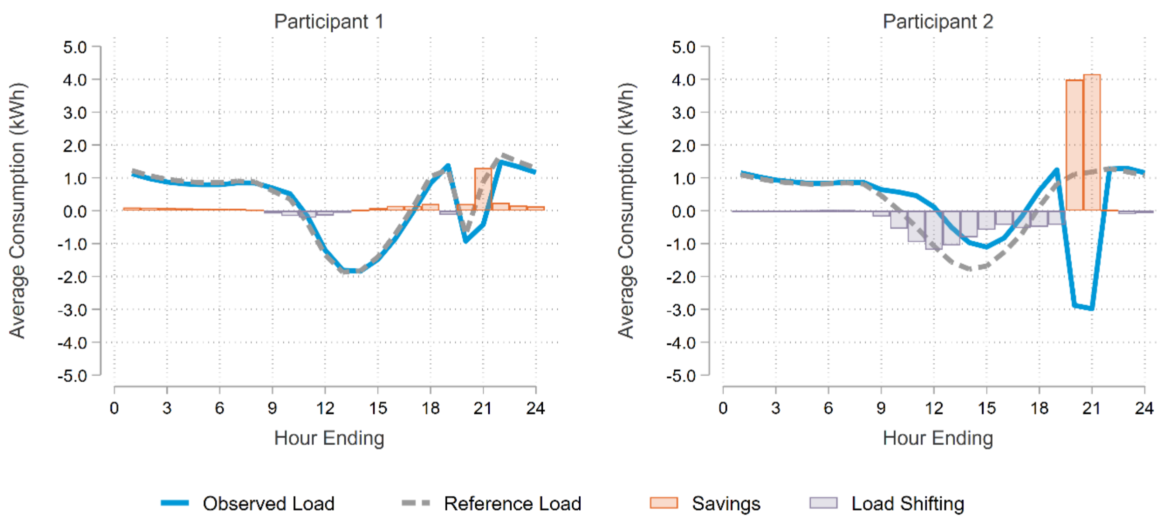
[98] *See Phase V Final Implementation Order* at 151-152.

not require significant modeling or manipulation as the counterfactual is zero discharge (the absence of a battery) and the inverter data does not contain noise from other end uses like AMI data from the EDC revenue meter. For each site, ECs should confirm whether the battery configuration is AC or DC-coupled to determine whether the battery telemetry data includes the effect of inverter losses and adjust for any losses accordingly. In addition, EDCs should evaluate the performance of rebated battery storage installations across both summer and winter seasons, regardless of when the battery was installed. If an EDC opts to have distinct operational modes (e.g., afternoon vs. morning and evening discharge), then strict customer agreements or contracts outlining the charge-discharge profile must be provided. In this case, the EC may not necessarily need to evaluate BTM battery storage across both seasons.

For BYOB daily load shifting, ECs should leverage pre-enrollment AMI to difference out any charging behavior that took place before enrollment in the daily load shifting program and isolate the incremental effect of the BYOB intervention. EDCs should require collection of the battery and solar adoption and interconnection dates for each participant in these storage-enabled daily load shifting programs. This is an important consideration so that the effect of the adoption of the battery itself (and solar) can be distinguished from the effect of the program. If the BYOB program design involves EDC agreements with battery manufacturers to aggregate demand reductions from their fleet of installed batteries, randomization should be explored. Most of the large and sophisticated battery vendors can optimize complex market signals so implementation of an on-off design with known performance hours will not pose a challenge.

Figure 18 illustrates the importance of including pre-program (but post-battery) data in the impact evaluation. The left panel (Participant 1) shows a reference load pattern based on pre-program behavior that includes some battery discharge in the evening. The right panel (Participant 2) shows a different reference load pattern with little to no evening battery discharge prior to program participation. To capture the program effect, the preexisting battery behavior should be used as a baseline so as not to bias impact measurement. For example, using the right panel baseline behavior as the counterfactual for the left panel observed load would artificially increase the measured impact, as the default discharge behavior would not be considered when measuring the program effect.

Figure 18: Measuring Impacts with Different Baseline Behaviors



It may be possible to collect pre-enrollment inverter data for a BYOB program where the battery was already installed. ECs are encouraged to collect pre-enrollment inverter data from vendors where possible to correctly capture the baseline pre-enrollment behavior of battery charge and discharge.

6.2.3.3 Smart Device Optimization

There are a wide range of technologies that can be used to achieve daily load shifting within residential homes, including smart thermostats, water heaters, plug load controllers, and other smart appliances. This section focuses on the use of smart thermostats and water heaters, though there are many other smart devices that can be used for daily load shifting.

Smart thermostats control two of the largest end-use loads in most homes (heating and cooling), making them a practical avenue for achieving daily load shifting in the residential sector. As a mass-market technology that manufacturers can manipulate remotely, they are well-suited to an alternating treatment design. For example, it should be straightforward to withhold a subset of participants from dispatch on any given day without affecting the dispatch of other participants and with minimal impacts to the customer. Since participants would typically expect their thermostat operation to be automatic and adjustments subtle, it is unlikely that they would interfere with thermostat control, much less notice that they are not being dispatched on a given day. This would mean the "control" and treated group are identical except in the operation of the thermostat itself.

If enrolled devices can be mapped to their EDC account, then AMI data should be used to estimate the impacts of smart device optimization. With AMI data there is no need to assume the average connected load of the equipment or prevalence of multi-zone systems. Thermostat runtime data also can be used instead of AMI data to evaluate smart thermostat optimization impacts if EDCs are unable to associate each device with a specific utility account. However, location data at least as granular as 5-digit zip code must be available to ECs at the device-level to ensure EDCs are only claiming impacts from within their service territory. The connected load assumption used to convert runtime to energy is a key assumption when device telemetry is used for impact analysis. The basis for all connected load assumptions should be documented up front in the EM&V plan and supported by Pennsylvania-specific research like the 2023 Pennsylvania Residential Baseline Study[99] to the extent possible.

Smart water heaters are another smart device option for daily load shifting and are similarly well-suited to the alternating treatment design. Smart water heaters can be "instructed" to optimize energy usage by preheating and storing domestic hot water during times in the day when capacity is more abundant, reducing the need for heating energy during peak hours. Since these devices can receive signals remotely, a similar design where a subset of water heaters is withheld from "optimization" could be leveraged to increase measurement accuracy for load shifting impacts. Table 40 shows the experimental design for Southern California Edison's SmartShift Rewards Hot Water program. For this daily load shifting program, participants experience the optimization algorithm for three weeks and then spend a week acting as controls each month.

Table 40: Experimental Design for SCE SmartShift Rewards Hot Water Program

Group	Four-Week Cycling Pattern			
	Week 1	Week 2	Week 3	Week 4
Group A	Load shifting in effect	Load shifting in effect	Load shifting in effect	No shifting
Group B	Load shifting in effect	Load shifting in effect	No shifting	Load shifting in effect
Group C	Load shifting in effect	No shifting	Load shifting in effect	Load shifting in effect
Group D	No shifting	Load shifting in effect	Load shifting in effect	Load shifting in effect

6.2.3.4 C&I Daily Load Shifting

Individually matching C&I accounts (especially Large C&I) with unique load profiles can be challenging for several reasons. Industrial load patterns are often a function of production schedules rather than outdoor temperature conditions, and evaluators may not have insight into the production schedules. Since the number of Large C&I customers operating in each area is much smaller than the number of residential or small commercial customers, the pool of potential matches is often more limited. In cases where suitable matched controls cannot be identified, non-participant information can still be leveraged using granular profiles modeled as synthetic controls. Granular profiles, which have been used for the evaluation of energy efficiency and load shifting offerings in California since the COVID-19 pandemic, represent aggregated load patterns of eligible customers, grouped by industry type. These granular profiles are publicly available through CALMAC,[100] and, when used as synthetic controls, show

[99] See *2023 Pennsylvania Residential Baseline Study* at: Weblink
[100] The methodology and development of granular profiles are documented by CALMAC at: Weblink.

comparable results to the use of matched controls without raising concerns about non-participant customer data privacy.

Constructing a synthetic control group involves adding one or more aggregated hourly control group profiles to the site-specific regression specification as an explanatory variable. This approach relies on these added profiles to construct a synthetic control that exploits the relationship of consumption patterns between the aggregated control group loads, and the participant loads during the pre-treatment period to predict participant usage during the post-treatment period. A low-effort alternative to designing and maintaining granular profiles for Act 129 purposes is to simply leverage the class profiles that are already published by the Pennsylvania EDCs,[101] for the benefit of Electric Generation Suppliers. Inclusion of a class load index in the model specification along with weather and temporal factors provides the accuracy benefits of non-participant information for sites that prove challenging to match individually.

Additionally, EDCs can employ a recruit-and-delay tactic when implementing a load shifting program for C&I accounts. This strategy would involve accepting C&I customers into the program, but not implementing any load shifting, until a baseline C&I profile can be collected. For example, an EDC may conduct a facility audit in May to identify improvements, and wait to implement lasting operational changes, such as shifting work schedules, until June. All the control days in this case are taken upfront between the audit and the intervention, and the EDC will be held harmless for this control period and credited with the average impact observed once the intervention began.

The length of Phase V presents a potential challenge for C&I daily load shifting assuming EDCs allow participants to remain enrolled in the program for consecutive program years. Most of the viable evaluation methods for C&I daily load shifting will leverage pre-treatment data. As the performance period becomes further separated in time from the pre-treatment period, the chances of operational changes unrelated to the program increase. Changes to operating hours, production levels, tenancy, expansions, or contractions could create a situation where the baseline model no longer accurately represents facility load absent the load shifting intervention. However, if the EDC "re-baselines" a facility, the load shifting actions the program induced could be embedded in the baseline and cause impacts to be underestimated. In general, EDCs should establish a single baseline period and use it for the duration of Phase V unless operational changes render that original mathematical model of facility energy use unreliable. Since load shifting is expected to alter the timing of electric consumption rather than the total volume, analysis of total energy consumption is a useful indicator for whether re-baselining should be considered. Possible scenarios that may warrant re-baselining include:

- If the average weekly or annualized consumption of the facility increases or decreases by more than 10%, and the change isn't explainable by weather or manufacturing production
- The primary business activity of the facility changes (e.g. warehouse space begins to be used for light manufacturing and assembly of products)

---

[101] PECO class profiles are published monthly at: Weblink.

- The total weekly operating hours change by more than 10% in either direction
- The facility installs solar PV or some other distributed energy resource that fundamentally alters the consumption profile of the facility
- The customer change tariffs in a way that fundamentally alters its billing determinants

ECs are encouraged to explore potential re-baselining cases with the SWE prior to updating the reference load model for a participating facility.

6.2.3.5 Time-Varying Pricing

Perhaps the most complex potential daily load shifting offering EDCs may elect to implement in Phase V of Act 129 is time-varying electric pricing. Electricity is widely accepted to be an elastic product, meaning consumer usage responds to changes in price. Decades of experimentation with time-varying rate structures have shown that customers will shift usage from higher price periods to lower price periods when facing a retail electric rate that varies across the day. There are virtually unlimited rate structures which an electric utility could implement to motivate customers to modify their consumption of electricity. Time differentiation could apply to the generation component of the rate (for customers who take their energy supply from the EDC) or to the distribution charges. It is outside the scope of this protocol to document the array of time-varying pricing structures, so we describe the concepts in general terms and focus on the measurement and verification considerations for the EDCs and their EM&V contractors.

The Phase V Final Implementation Order addresses the ability of EDCs to claim peak demand savings from time-varying pricing.[102] Provided the approved EE&C plan includes customer education, targeted behavioral programming, or monetary incentives to accelerate TOU adoption and bolster load impacts, the EDC can claim Act 129 savings from all the new enrollments. There is no need to separate uplift from natural enrollment since compliance targets are based on gross verified savings. The extent to which Act 129 support increased TOU rate adoption would be a useful topic to explore in the net impact evaluation.

The way enrollment into the time-varying rate is designed is a key consideration when deciding on an evaluation method. Often, customers are defaulted onto time-varying rates, meaning a group of customers is switched onto that rate unless they opt out. Where possible, withholding a randomized control group entirely from enrollment into the rate provides the best opportunity to measure default TOU impacts. Using customers who opt out of a default TOU rate is not an advisable control group strategy as deniers and compliers are likely to have important structural differences.

In contrast to the default TOU design, customers are sometimes given the option to opt in to the TOU rate. With opt-in TOU, where customers self-select into treatment, an RCT design is not possible. One possible way to incorporate randomization into the assignment of treatment would be with the RED design, where customers are randomly assigned whether they will be encouraged to opt in to the TOU

---

[102] *Phase V Final Implementation Order*, page 123.

rate. If the RED design is not feasible, a matched control group is appropriate. Since TOU opt-in rates are typically low, there should be a large pool of customers to construct a matched control group from.

The alternating treatment design is likely not appropriate for either default or opt-in TOU. Like EV managed charging offerings, where the response to the intervention likely becomes ingrained over time among participants, there is no way to easily "switch off" the price response even if some are withheld from the rate on a given control day. Additionally, this would add complexity to EDC billing processes. Participants may not remember their previous behavior in the absence of the TOU rate and may simply behave as they have been "coached" if education was part of the intervention. This makes the alternating treatment design infeasible for the purpose of measuring TOU impacts.

With default or opt-in TOU, the rate is in place every day after enrollment. Therefore, if a matched control group is used, all matching should be done on pre-enrollment data. In cases where there is not sufficient baseline consumption data before TOU enrollment, for example in cases where TOU enrollment happens right after account creation, it is acceptable to drop these accounts from the analysis and scale up impacts to the whole population.

Program design as well as the feasibility of randomization should be the primary decision factors in whether an RCT, RED, or matched control group is used. Applying price elasticities or per-participant impact assumptions should be avoided, except if those assumptions are taken from a previous ex post evaluation of the same rate from a previous year of Phase V. Section 5.2.1 of the California Load Impact Protocols[103] provides additional guidance for the evaluation of non-event based resources like time-varying rates and methods that mitigate selection bias where the randomization of treatment is not possible.

6.2.3.6 Behavioral Messaging

In Phase III of Act 129, several EDCs implemented Behavioral Demand Response (BDR) programs in which participants were notified via text, email, or phone call of peak conditions and encouraged them to reduce their electric demand during a specific set of hours. Peak Time Rebate (PTR) programs are a common offering in other jurisdictions. PTR programs are similar to BDR with the addition of financial incentives for the estimated load reduction. Typically, settlement occurs at the individual customer level, which means the performance estimates are inherently noisy. Participants are compensated for estimated reductions but held harmless for estimated increases. This payment asymmetry can make PTR programs economically inefficient.

For Phase V of Act 129, PPL plans to pilot daily load shifting PTR. Instead of being encouraged to conserve electricity on specific event days, the messaging will focus on general suggestions regarding when it is most and least efficient to use electricity during the summer and winter. The periodic communications could highlight participant economics (lowering PLC or TOU charges), greenhouse gas avoidance (using energy when the grid is cleaner), load factor improvement, or alleviating grid

---

[103] See California Load Impact Protocols. Weblink. Section 5.2.1.

constraints that drive costs for all customers. Gross verified impacts will be measured via a difference-in-differences calculation between participants and a control group of non-participants. Ideally, the treatment and control groups would be randomly assigned. If random assignment is infeasible because enrollment is on an opt-in basis, a matched comparison group is a suitable alternative. Both treatment and control group homes should have AMI data available for the pre-PTR summer and winter to facilitate the difference-in-differences calculation.

While the payment asymmetry is an economic challenge for PTR, it does not pose an impact evaluation issue because evaluators do not require individual customer estimates. The objectives of the impact evaluation are the average and aggregate treatment effects. Noise reduction is a key benefit of aggregation and panel regression methods. The average effect across thousands of homes is far easier to measure precisely than the impact in a single home. Daily PTR should be less noisy than event-based PTR due to aggregation across the many performance days in a season. ECs should compare the gross verified peak demand savings to the sum of individual customer settlement estimates (if applicable) and discuss the economic implications of their findings.

6.3 M&V Guidance for Large Non-Residential Solar PV Systems

This section of the Framework provides guidance to the EDCs and their evaluators on how to conduct solar photovoltaic (PV) on-site M&V for large systems that exceed the 2,000 MWh of annual reported generation threshold as defined in measure 3.11.6 of the 2026 TRM.

6.3.1 Measure Eligibility

The goal of Act 129 EE&C plans is to reduce consumption and congestion on the state's electric system. Projects that generate savings for Act 129 programs must offset existing facility loads. Virtual metering of multiple sites, within the guidelines provided by 52 Pa. Code § 75.14.(e),[104] will be permitted to define existing facility loads. New construction facilities can utilize energy models to estimate electricity consumption. All projects must receive an interconnection agreement and comply with all EDC requirements to be eligible. If the annual generation of a solar PV project exceeds consumption of the most recent (or forecasted) 12 months, the EDC evaluator should provide evidence-based justification that the excess generation will be consumed by the same customer within the same distribution feeder. Otherwise, claimed generation (kWh) should be capped at existing facility load. EDCs should provide their evaluators with historic facility consumption data for solar participants to facilitate this comparison.

6.3.2 TRM Methodology Summary

Non-residential solar PV installations with an annual reported generation greater than 2,000 MWh require more rigorous M&V because of their larger portfolio weight while smaller systems can utilize a streamlined M&V approach. Under the TRM guidance, systems below the threshold can apply PVWatts

[104] 52 Pa. Code § 75.14.(e). Weblink

modeling, based on array-specific data, to estimate annual generation and demand savings. Systems over the threshold are required to incorporate customer-specific generation data. This data can be metered directly off the system, provided through inverter tracking data, or collected from another reliable system source.

6.3.3 Metering and Modeling Considerations

The production levels of solar PV systems are dependent on variables that can be categorized into three groups,

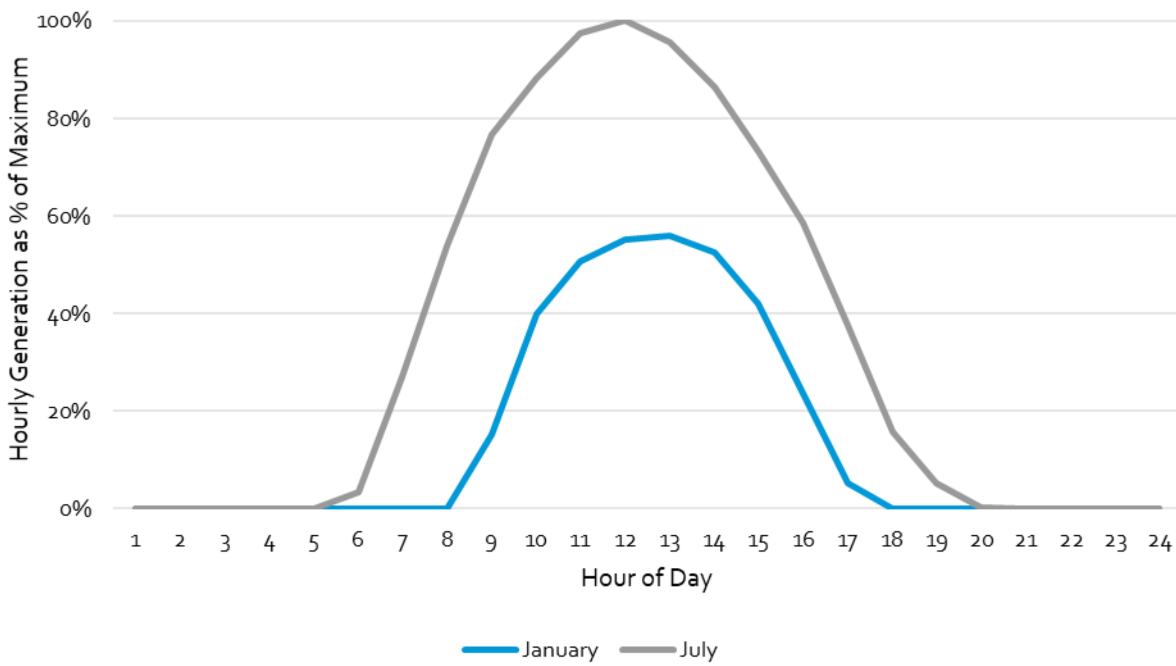
- Static variables
- Cyclical variables
- Weather-dependent variables

Static variables are fixed once the system design and construction are complete. This includes installed capacity, array tilt, and azimuth. Cyclical variables change across the year and follow predictable patterns, like sunrise/sunset times, solar elevation, or the presence of a tracking system. Weather-dependent variables introduce the greatest uncertainty to solar PV production. Solar irradiance influenced by cloud cover and precipitation can drive large generation shifts across the day. Snow coverage or frost blocks solar rays from reaching the array and can reduce generation to near zero, even on sunny days. Ambient temperature will influence solar panel generation efficiency. Each of these variables influence the solar irradiance that reaches the system, making measurement of the irradiance data point a high priority for all projects.

These large solar PV systems will contribute substantial energy and demand savings toward Act 129 compliance targets. Therefore, it is important that site-collected data informs actual system performance for both energy and peak demand metrics. Figure 19 illustrates the average hourly generation profile for a southeast-facing solar PV system in the months of January and July.[105]

[105] The SWE notes that the 8760 outputs from PVWatts and SAM do not adjust for daylight savings time. Due to this the hours identified for summer peak demand should shift back one hour to model system performance in the 2pm to 6pm EDT period.

Figure 19: Daily Average Solar PV Generation Profile Across Seasons



Summer peak demand savings in Figure 19 occur between the hours ending 15 and 18 in the July example, when daily production is decreasing. Production decreases during periods with fewer daylight hours and energy generation and associated demand savings will go to zero in the later hours associated with summer peaks. Generation data from the beginning and end of each day will have increased uncertainty when compared to hours with greater expected production, thus increasing the importance of summertime data collection for any performance models. Direct measurement of generation during the summer peak demand period will be more consistent with the daily solar profile and seasonal weather patterns that occur during those months, helping to reduce uncertainty in peak demand savings estimates. Solar production during the Act 129 winter peak demand definition is expected to be small, but non-zero.

6.3.4 Data Collection Guidance

For systems that exceed the threshold the 2,000 MWh annual reported generation analysis should incorporate a minimum 90 days of site generated data. The SWE prefers the use of generation data across the full summer peak demand period but acknowledges that the end of program year and reporting deadlines create practical challenges for data collection in June, July, and August. As a result, a minimum of 30 days of data collection during the summer peak period will be required to inform peak demand savings. This can be paired with 60 days of non-summer peak period data collection to meet the 90 day requirement. Site-specific solar irradiance data is preferred and can reduce performance uncertainty, but actual weather data sets from a nearby Class A weather station can be substituted, if needed. Energy generation models developed from on-site metering need to incorporate weather normalization to minimize the influence of weather-dependent variables and changes to solar irradiance that are inherent to site-collected generation data. TMY3 and TMYx are acceptable weather standards. Modeling for solar PV systems can also apply the National Solar Radiation Database (NSRDB) to align with PVWatts and leverage improved site resolution, or TMY2 data if snow depth is applied as a regression variable. Savings estimates provided by the evaluator developed using models which incorporate on-site data shall be considered ex post savings. The SWE expects most analyses will apply irradiance measurements from nearby weather stations, NSRDB, Solcast, or Aurora Solar in their models.

The SWE requires each evaluator to design a methodology for estimating ex post savings for large solar PV projects that meet these requirements. SSMVPs are required for all projects per this Evaluation Framework. In lieu of reviewing all project SSMVPs the SWE will allow each evaluator to provide a generic solar PV project M&V plan or calculator that will be followed for all large solar projects and that summarizes:

- The duration and expected time frame for on-site data collection
- The source(s) for solar PV generation data
- How irradiance data during the generation period will be measured or estimated
- How generation and irradiance data will be used to develop annual generation models for the solar PV systems under evaluation
- How the models developed in the prior steps will be extrapolated to provide weather-normalized Act 129 energy and summer peak demand savings

6.4 Conservation Voltage Reduction Program Evaluation

The Phase V Final Implementation Order specifies that FTM measures are eligible, provided they represent less than 10% of the EDC's compliance savings. Stakeholder comments and the Commission's disposition specifically addressed Conservation Voltage Reduction (CVR) as a potential low-cost, high-yield FTM measure. PECO and FirstEnergy implemented CVR in Phase I and PPL indicated in their Phase V EE&C Plan their intention to allot up to $1 million to implement a CVR Pilot. CVR programs target reductions in energy consumption and peak demand by decreasing the voltage of the distribution system before power is delivered to the customer. Typically, this intervention involves either manually or remotely controlling line tap changers and downline voltage regulators to lower the regulated voltage band in front of customer meters.

CVR programs can be implemented by either permanently lowering the voltage of the distribution system ("once on, always on") or through an event-based system. In the case of an event-based system, voltage reduction is only applied to the distribution system on certain days or at certain times of the day to achieve targeted reductions in energy usage and peak demand. Each of these approaches can have benefits and drawbacks for both program operations as well as after-the-fact evaluation.

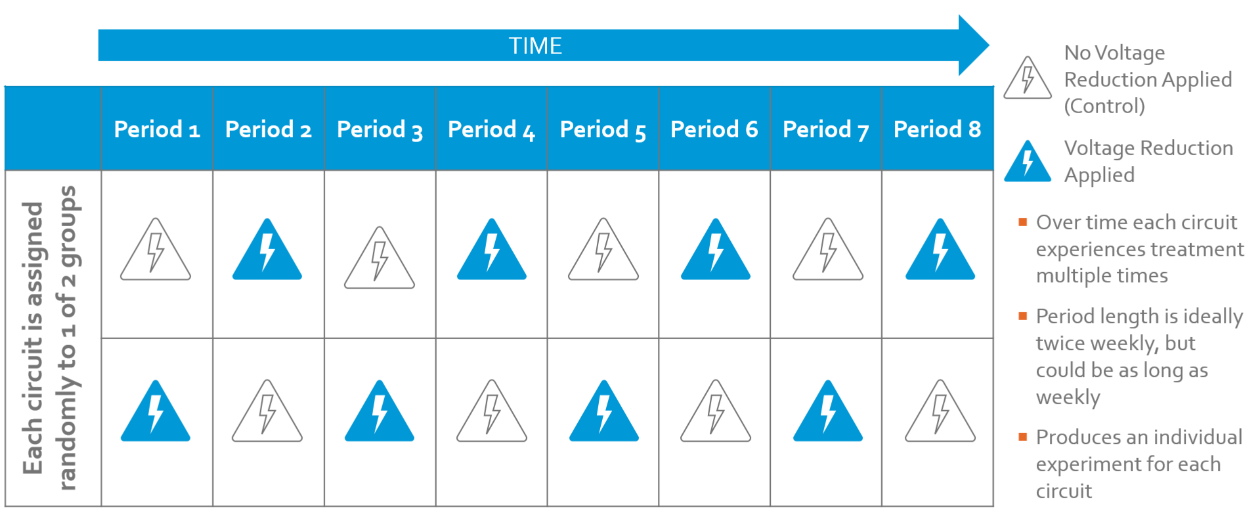
6.4.1 Evaluation Planning

CVR programs typically only target a subset of the circuits within the distribution system and seek to reduce energy consumption by around 1% on treated circuits. The ability to accurately estimate energy and peak demand reduction from CVR is contingent upon a sound experimental design that systematically eliminates alternative explanations for changes in energy use other than CVR. Ex post, or after-the-fact analysis, relies heavily on econometric modeling of energy demand to develop a counterfactual of what energy demand would have been in the absence of using CVR technology.

6.4.1.1 Pilot Experimental Design

The experimental design of any CVR pilot is crucial to accurately detecting the impacts of the intervention. Evaluators should choose a design that maximizes the amount of explanatory power for the best opportunity to rule out all other explanations for changes in loads other than the program intervention. The standard approach to conducting research on CVR is to use a rotating treatment framework.[106] In this framework, a set of distribution circuits is chosen to be treated and over the course of the study, ideally for a full year, and each circuit rotates the voltage reduction treatment on and off at set intervals. A rotating treatment design allows each treated circuit to be observed in the treated and untreated state throughout the experiment and makes use of controls in each period, as at least one circuit is not treated. Figure 20 shows an example of a rotating treatment design for a CVR program, where each circuit is observed in both the treated and untreated state across the study. The SWE Team recommends that any CVR pilot programs in Pha se V of Act 129 be implemented using a version of this experimental design.

Figure 20: Example Rotating Experimental Treatment Design for CVR



The example design in Figure 20 shows a program with only 8 periods, but that number is only meant to be illustrative. Practical constraints on the number of groups to track and the frequency at which

106 See IESO's Evaluation, Measurement and Verification Protocols and Requirements V3.0, Volume III. Weblink

operations teams can apply the voltage reductions will limit the number of circuit groups and period lengths implemented. The SWE recommends at least two groups so that each day of the study includes both "on" and "off" information as it will ensure a balanced data set. The CVR treatment should be alternately at least weekly for nine months over the summer, shoulder and winter seasons. Observing the effects of the treatment in each season is ideal, as it will determine if there is any seasonality to the effects of CVR.

While the SWE Team recommends the above experimental design to maximize the effectiveness of the study, we recognize that operational constraints may restrict a pilot's ability to implement a study design that requires frequent modification of equipment. In the case that an EDC plans to implement a CVR study using a "once on, always on" approach, where voltage reduction is applied to a set of circuits for the full duration of the study period, the most robust quasi-experimental method available is a difference-in-differences model with a group of matched control circuits. This experimental design approach makes use of non-treated circuits to incorporate information on general energy use trends into the model. In the case of CVR, though, this approach is challenging, because there is only a relatively small number of circuits within each EDC's territory, which means that finding a good match for each treated circuit is challenging. If the matched control group is not a good proxy for the behavior of the treated circuits without intervention, then evaluators will struggle to produce precise and unbiased estimates of the treatment effect.

6.4.1.2 Power Analysis

Regardless of the experimental design method chosen, the evaluation team should implement a power analysis prior to study implementation to test the expected precision of results. This test should be implemented as a "false experiment" where Supervisory Control and Data Acquisition (SCADA) or meter data is collected for all the circuits in an EDC territory for a full 365 days and each experimental design is applied to the data where no treatment was ever in place. This means that, by definition, the treatment effect is 0 and any treatment effect estimated through modeling is strictly model error. The power analysis should test any potential variables that can be controlled by evaluators when designing the experiment.

In the case of the recommended alternating treatment design, the two major variables that can be controlled are the sample size, or the number of circuits where CVR will be implemented, and length of the period of rotation, or the number of days between turning treatment on and off. Figure 21 summarizes the power analysis simulation process, which can be used to quantify the explanatory power of each potential design.

Figure 21: Power Simulation Example

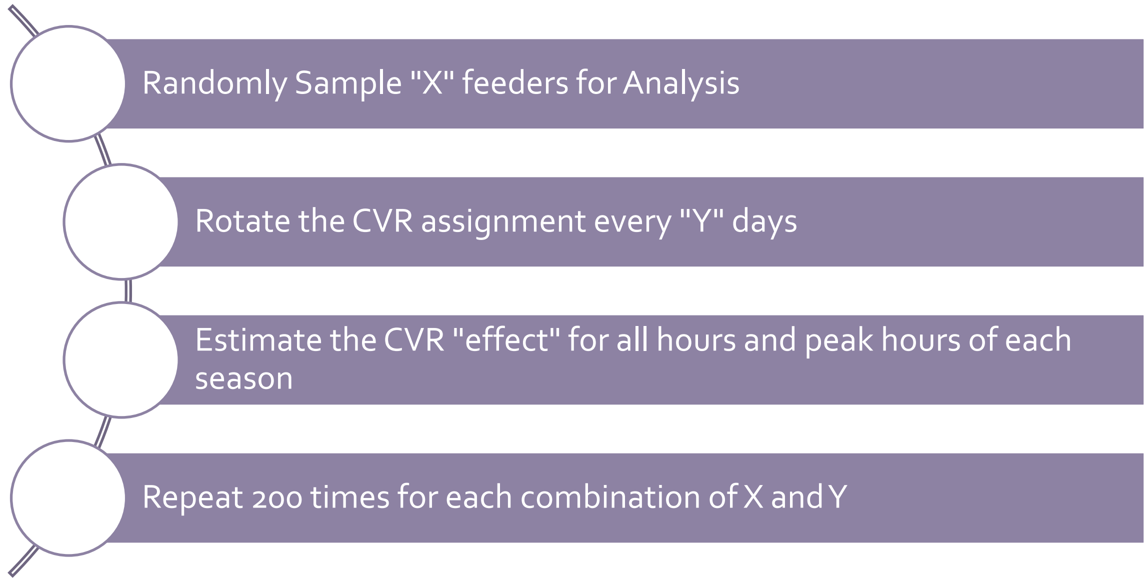
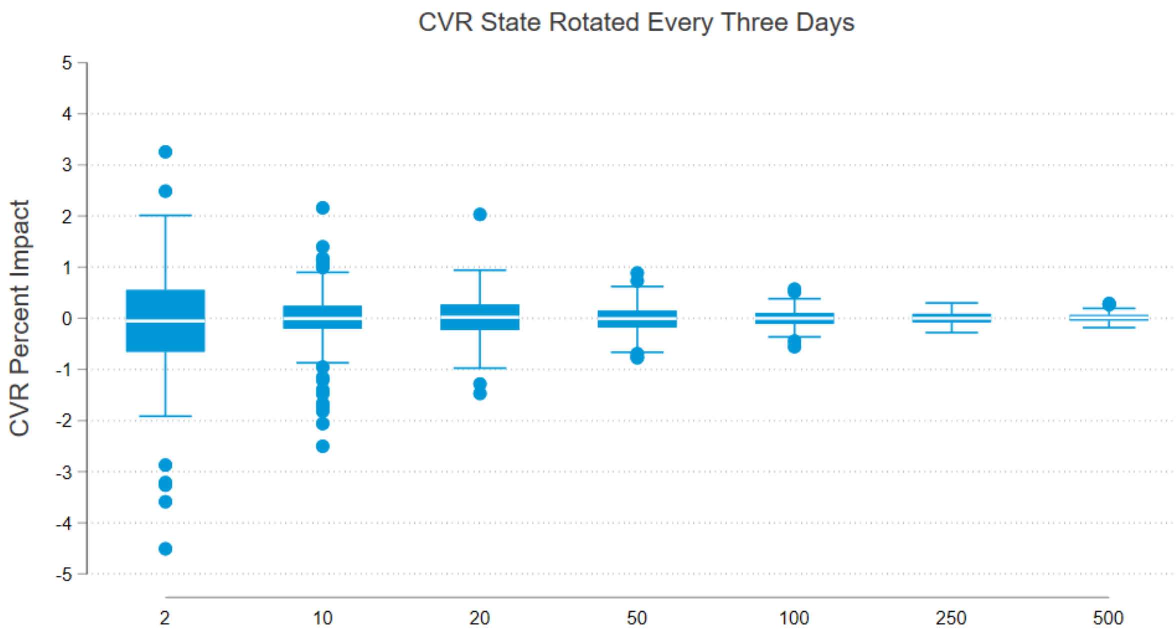


Figure 22 shows example power simulation results that assume the CVR treatment is rotated every three days for the sampled feeders. The "box-and-whisker" plot shows the distribution of results across the 200 iterations for each sample size. Even with small sample sizes, the results of the 200 iterations are centered at zero, which is the true effect for this data. However, with small sample sizes, individual simulations can produce erratic answers, highlighting that larger sample sizes lead to more precise results. The distribution of errors from the power simulation will approximate the modeling error evaluators can expect under various designs.

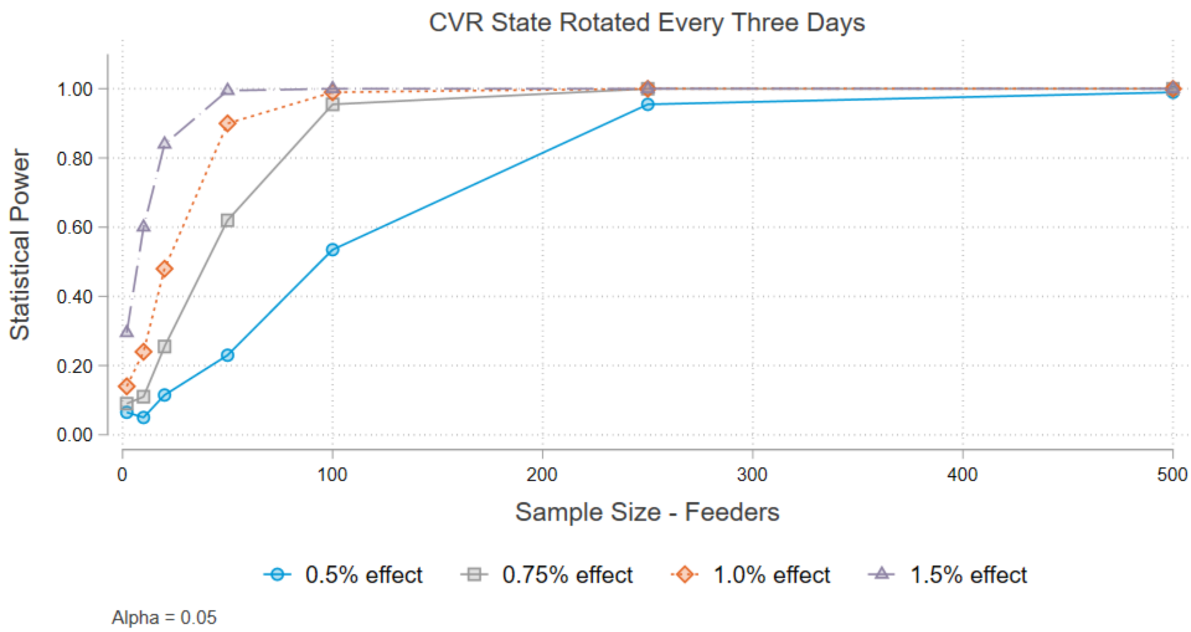
Figure 22: Example Distribution of Estimated Effects by Sample Size



The outputs of the simulation will give evaluators an idea of the expected distribution of errors given different sample sizes and treatment rotation periods, which can then be compared against the expected size of the treatment effect from CVR to determine if there is enough explanatory power to differentiate the treatment from residual noise in the data. Industry studies have consistently found reductions in energy consumption of approximately 0.8% for every 1% reduction in service voltage.[107] The final experimental design chosen should be able to detect a change of this magnitude given the EDC's expected voltage adjustment for both energy savings and peak demand reductions from CVR. Figure 23 shows example statistical power curves assuming the CVR treatment is cycled every three days. Statistical power is the likelihood of detecting an effect if one exists. If the true effect is 1.5%, in this example, a study of 50 feeders is virtually guaranteed to produce statistically significant results. If the true effect is only 0.5%, however, a sample size of 250 feeders would be needed.

[107] See Costs and Benefits of Conservation Voltage Reduction. NRECA-DOE Demonstration Project, 2014. Pg. 2. Weblink

Figure 23: Example Statistical Power Curves by Sample Size and Effect Size



The results of a power analysis conducted by evaluators should be used to determine the final implementation of the study design. The final design should balance the need for statistical power with the operational and budget limitations of the CVR deployment.

6.4.2 Impact Estimation

6.4.2.1 Data Requirements

To measure impacts from a CVR program, evaluators can base their analysis on data measurements from two different sources. First, because CVR is applied for an entire distribution circuit, measurements of power demand at the substation or circuit feeder can be used to model impacts from treatment. Using distribution SCADA data comes with the advantage of having direct measurements of circuit voltage. Second, customer meter data for all customers on all circuits included in the study can be aggregated together at the circuit level to form the basis of the analysis. Unless individual customer meters all have hourly voltage readings, the AMI data will likely need to be supplemented with distribution circuit-level readings of voltages by hour.

The only other data sets required for modeling are weather data, to allow for outdoor temperature controls to be included in the model, and information on when CVR was implemented and on which circuits. Despite the fact that the alternating treatment design means that CVR treatment should not be correlated with outdoor temperatures, and therefore loads, random chance can mean that outdoor temperatures may differ during treatment periods from non-treatment periods. In this case, the inclusion of weather data in the model can help explain any differences in loads due to differences in experienced weather.

6.4.2.2 Impact Modeling

Final impact modeling from implementation of CVR should first and foremost quantify the estimated energy and peak demand savings by season over the course of the study. Primary results from the evaluation should focus on quantifying the extent of the treatment effect of lowering voltage on a circuit. Table 41 shows the SWE's minimum expectation of final regression modeling results from CVR pilot studies. The expected independent and treatment variables assume the use of the alternating treatment design. If a DiD approach is instead chosen, then final modeling will need to include a post period indicator alongside the treatment indicator. Each row represents a model that should be fit to characterize the effect that the treatment had on loads.

Table 41: Expected Modeling Approach and Impact Estimates

Model Type	Dependent Variable	Independent Variables	Treatment Variable	Output Interpretation
Level-Level	MW (or MVA)	Weather and indicators for hour and day of week	CVR on/off (0,1)	Average change in MW (or MVA) when CVR is in effect
Log-Level	Natural Log of MW (or MVA)	Weather and indicators for hour and day of week	CVR on/off (0,1)	Percent Change in MW (or MVA) when CVR is in effect
Log-Log	Natural Log of MW (or MVA)	Weather and indicators for hour and day of week	Natural Log of Voltage	Percent Change in MW (or MVA) per 1% change in Voltage

Equation 20 shows an example regression specification that could be used to determine the ATE from CVR (β_c). The treatment effect can be interpreted as the average change in load (MW) for a treated circuit when CVR is in effect. The β_c coefficient from Equation 20 should be negative, indicating that lowering voltage reduces power demand on average, but the expected magnitude of the coefficient depends on the average size of the treated circuits. This model corresponds to the level-level of Table 41.

Equation 20: Example Model Specification

$$MW_{f,t} = \beta_0 + \beta_c * CVR + \beta_h * hour + \beta_{h,t} * hour * temp + \beta_t * temp + \beta_d * day\ of\ week + \epsilon_{f,t}$$

Table 42 defines the terms from Equation 20.

Table 42: Example Model Definition of Terms

Variable	Definition
$MW_{f,t}$	measured demand on circuit f at time t.
β_0	Intercept of the regression equation.
β_c	Treatment effect. The coefficient on the CVR indicator variable.
CVR	Indicator variable equal to 1 when CVR is in effect, 0 otherwise.
β_h	The array of coefficients for each hour of the day
$hour$	Indicator variable for the hour of the day
$\beta_{h,t}$	The array of coefficients associated with each hour of the day for temperature

Variable	Definition
$temp$	The outdoor temperature of a nearby weather station
β_t	The coefficient for outdoor temperature
β_d	The array of coefficients for each day of the week
$day\ of\ week$	Indicator variable for each day of the week
$\epsilon_{f,t}$	The error term.

The log-level model described in Table 41 only differs from the level-level model in that the dependent variable, MW, has been transformed by taking the natural log of MW. This means that the estimated coefficient for the treatment variable can be interpreted as the percentage change in power demand when CVR is in effect. This estimate is useful because it is no longer dependent on the average size of the circuits that were a part of the study. This means that the estimate of the effect from CVR could more easily be applied to other circuits outside of the study regardless of their size, assuming that the expected change in voltage was like those observed in the study.

Results from the log-log model are even more generalizable and can also be applied to other circuits where CVR is being applied, even those that change the voltage on the circuit by a different amount than the average that was observed in the study. The log-log model is specified in the same way as the log-level model, except that it no longer includes a CVR indicator and rather includes the natural log of voltage observed during the study. The coefficient on this model variable can be interpreted as the percent change in power demand for a 1% change in voltage. This means that savings from a CVR implementation that changes voltage by a different amount than the average that was observed in the study can be estimated using this value. This coefficient is referred to as the CVR factor.

6.4.3 Annualization and Reporting

6.4.3.1 Data Requirements

Energy and peak demand savings from the implementation of CVR on a circuit are expected to have an EUL of 15 years. To claim the savings determined after the evaluation of the measurement period, the EDCs and their evaluators must annualize the expected savings over a 15-year period. Annualization of results will require three different data sets:

- CVR factor from the results of the log-log model.
- Average annual expected voltage reduction percentage by feeder or substation where CVR will be implemented.
- Average annual weather-normalized forecasted energy sales and seasonal peak load by feeder or substation where CVR will be implemented.

6.4.3.2 Calculation of Gross Verified Savings

Final verified energy savings claims from CVR should be calculated separately for each feeder or substation where CVR will be implemented. Calculating savings separately by circuit will allow for variable percent changes in voltage to be used, based on each circuit's characteristics and operational constraints. Equation 21 shows the equations that should be used to calculate annualized savings from CVR. The peak demand savings should be calculated separately for both summer and winter demand.

Equation 21: CVR Annualized Savings Calculations

Energy:

$$\begin{aligned} &Verified\ Savings \\ &\quad = CVR\ Factor * \%\ Change\ in\ Voltage \\ &\quad * Weather\ Normlaized\ Average\ Annual\ Energy\ Sales \end{aligned}$$

Peak Demand:

$$\begin{aligned} &Verified\ Savings \\ &\quad = CVR\ Factor * \%\ Change\ in\ Voltage \\ &\quad * Weather\ Normlaized\ Average\ Annual\ Seasonal\ Peak\ Demand \end{aligned}$$

The calculations shown in Equation 21 assume that evaluators only estimate a single CVR factor during the impact estimation from the pilot activities. If the evaluator determines that there is enough available data to either estimate separate factors by season or a separate factor for changes in demand during the peak period, then those factors should be used where appropriate. Separate CVR factors for each of the time periods mentioned would capture any potential differences in the load response to decreased voltage based on the time of day or time of year. While separate estimates of CVR factors could potentially lead to a more accurate picture of savings, uncertainty in the feasibly of estimating separate factors means that EDCs should only be required to estimate an overall CVR factor at a minimum

7 Final Remarks

The primary objective of the EDC EE&C programs is to reach the level of savings specified in Act 129 in a meaningful, efficient, and cost-effective manner. It is the desire of the SWE to work closely and collaboratively with the PUC and EDCs to develop and implement an evaluation and audit process that will produce significant and standardized impact results, at the lowest cost, so that more funds may be allocated to customer-centric savings activities. The SWE must ensure that the evaluations are accurate and represent the actual impacts of the EE&C program with a targeted level of precision.

This Evaluation Framework outlines the expected metrics, methodologies, and guidelines for measuring program performance, and details the processes that should be used to evaluate the programs sponsored by the EDCs throughout the state. It also sets the stage for discussions among the EDCs, their ECs, the SWE Team, and the PUC. These discussions will help clarify the TRM, add new prescriptive measures to the TRM, and define acceptable measurement protocols for implementing custom measures to mitigate risks to the EDCs. The common goal requires that kWh/year and kW/year savings be clearly defined, auditable, replicable, and grounded in sound engineering estimates.

Appendix A Glossary of Terms

ACCURACY: An indication of how close a value is to the true value of the quantity in question. The term could also be used in reference to a model or a set of measured data, or to describe a measuring instrument's capability.

BASELINE DATA: The measurements and facts describing equipment, facility operations, and/or conditions during the baseline period. This includes energy use or demand and parameters of facility operation that govern energy use or demand.

BENEFIT/COST RATIO (B/C RATIO): The mathematical relationship between the benefits and costs associated with the implementation of energy efficiency measures, programs, practices, or emission reductions. The benefits and costs are typically expressed in dollars.

BIAS: The extent to which a measurement or a sampling or analytic method systematically underestimates or overestimates a value.

BILLING DATA: The term billing data has multiple meanings: (1) Metered data obtained from the electric or gas meter used to bill the customer for energy used in a particular billing period. Meters used for this purpose typically conform to regulatory standards established for each customer class. (2) Data representing the bills customers receive from the energy provider and used to describe the customer billing and payment streams associated with customer accounts. This term is used to describe consumption, demand, and account billing and payment information.

BUILDING ENERGY SIMULATION MODEL: A building energy simulation model combines building characteristic data and weather data to calculate energy flows. While hourly models calculate energy consumption at a high frequency, non-hourly models may use simplified monthly or annual degree-day or degree-hour methods.

CAPACITY: The amount of electric power for which a generating unit, generating station, or other electrical apparatus is rated by either the user or manufacturer. The term also refers to the total volume of natural gas that can flow through a pipeline over a given amount of time, considering such factors as compression and pipeline size.

COEFFICIENT OF VARIATION: The sample standard deviation divided by the sample mean (C_v = σ/μ).

CONFIDENCE: An indication of how close a value is to the true value of the quantity in question. A confidence interval is a range of values that is believed – with some stated level of confidence – to contain the true population quantity. The confidence level is the probability that the interval actually contains the target quantity. The confidence level is fixed for a given study (typically at 90% for energy efficiency evaluations).

CONSERVATION: Steps taken to cause less energy to be used than would otherwise be the case. These steps may involve improved efficiency, avoidance of waste, and reduced consumption. Related activities include installing equipment (such as a computer to ensure efficient energy use), modifying equipment (such as making a boiler more efficient), adding insulation, and changing behavior patterns.

CONSERVATION SERVICE PROVIDER (CSP): A person, company, partnership, corporation, association, or other entity selected by the EDC and any subcontractor that is retained by an aforesaid entity to contract for and administer energy efficiency programs under Act 129.

COST-EFFECTIVENESS: An indicator of the relative performance or economic attractiveness of any energy efficiency investment or practice when compared to the costs of energy produced and delivered in the absence of such an investment. In the energy efficiency field, the term refers to the present value of the estimated benefits produced by an energy efficiency program as compared to the estimated total program costs, from the perspective of either society as a whole or of individual customers, to determine if the proposed investment or measure is desirable from a variety of perspectives, such as whether the estimated benefits exceed the estimated costs.

CUSTOMER: Any person or entity responsible for payment of an electric and/or gas bill and with an active meter serviced by a utility company.

CUSTOMER INFORMATION: Non-public information and data specific to a utility customer that the utility acquired or developed in the course of its provision of utility services.

C_v: See Coefficient of Variation.

DEEMED SAVINGS: TRMs provide deemed savings values that represent approved estimates of energy and demand savings. These savings are based on a regional average for the population of participants; however, they are not savings for a particular installation.

DEMAND: The time rate of energy flow. Demand usually refers to electric power and is measured in kW (equal to kWh/h) but can also refer to natural gas, usually as Btu/hr, kBtu/hr, therms/day, or ccf/day.

DEMAND RESPONSE (DR): The reduction of consumer energy use at times of peak use in order to help system reliability, reflect market conditions and pricing, or support infrastructure optimization or deferral of additional infrastructure. DR programs may include contractually obligated or voluntary curtailment, DLC, and pricing strategies.

DEMAND SAVINGS: The reduction in the demand from the pre-retrofit baseline to the post-retrofit demand once independent variables (such as weather or occupancy) have been adjusted for. This term usually is applied to billing demand to calculate cost savings, or to peak demand for equipment sizing purposes.

DEMAND SIDE MANAGEMENT (DSM): The methods used to manage energy demand, including energy efficiency, load management, fuel substitution, and load building.

EFFICIENCY: The ratio of the useful energy delivered by a dynamic system (such as a machine, engine, or motor) to the energy supplied to it over the same period or cycle of operation. The ratio is usually determined under specific test conditions.

END-USE CATEGORY (GROUPS): Refers to a broad category of related measures. Examples of end-use categories include refrigeration, food service, HVAC, appliances, building envelope, and lighting.

END-USE SUBCATEGORY: This is a narrower grouping of measure types within an end-use category. Examples of end-use subcategories include lighting controls, LEDs, linear fluorescents, air-source heat pumps, refrigerators/freezers, central air conditioning, and room air conditioning.

ENERGY CONSUMPTION: The amount of energy consumed in the form in which it is acquired by the user. The term excludes electrical generation and distribution losses.

ENERGY COST: The total cost of energy, including base charges, demand charges, customer charges, power factor charges, and miscellaneous charges.

ENERGY EFFICIENCY: Applied to the use of less energy to perform the same function, and programs designed to use energy more efficiently. For the purpose of this Evaluation Framework, energy efficiency programs are distinguished from DSM programs in that the latter are utility-sponsored and -financed, while the former is a broader term not limited to any particular sponsor or funding source. *Energy conservation* is a related term, but it has the connotation of "doing without in order to save energy" rather than "using less energy to perform the same function"; it is used less frequently today. Many people use these terms interchangeably.

ENERGY EFFICIENCY AND CONSERVATION PLAN AND PROGRAM (EE&C): EE&C plan and program for each EDC in Pennsylvania.

ENERGY EFFICIENCY MEASURE: A set of actions and/or equipment changes that result in reduced energy use – compared to standard or existing practices – while maintaining the same or improved service levels.

ENERGY MANAGEMENT SYSTEM (EMS): A control system (often computerized) designed to regulate the energy consumption of a building by controlling the operation of energy-consuming systems, such as those for space HVAC, lighting, and water heating.

ENERGY SAVINGS: The reduction in use of energy from the pre-retrofit baseline to the post-retrofit energy use, once independent variables (such as weather or occupancy) have been adjusted for.

ENGINEERING APPROACHES: Methods using engineering algorithms or models to estimate energy and/or demand use.

ENGINEERING MODEL: Engineering equations used to calculate energy usage and savings. These models usually are based on a quantitative description of physical processes that transform delivered energy into useful work, such as heating, lighting, or driving motors. In practice, these models may be reduced to simple equations in spreadsheets that calculate energy usage or savings as a function of measurable attributes of customers, facilities, or equipment (e.g., lighting use = watts × hours of use).

EVALUATION: The performance of studies and activities aimed at determining the effects of a program; any of a wide range of assessment activities associated with understanding or documenting program performance or potential performance, assessing program or program-related markets and market operations; any of a wide range of evaluative efforts including assessing program-induced changes in energy efficiency markets, levels of demand or energy savings, and program cost-effectiveness.

EVALUATION CONTRACTOR (EC): Contractor retained by an EDC to evaluate a specific EE&C program and generate ex post savings values for efficiency measures.

EX ANTE SAVINGS ESTIMATE: The savings values calculated by the program ICSP, stored in the program tracking system and summed to estimate the gross reported impact of a program. Ex ante is taken from the Latin for "beforehand."

EX POST SAVINGS ESTIMATE: Savings estimates reported by the independent evaluator after the energy impact evaluation and the associated M&V efforts have been completed. Ex post is taken from the Latin for "from something done afterward."

FREE-DRIVER: A non-participant who adopted a particular efficiency measure or practice as a result of a utility program but who did not receive a financial incentive from a Pennsylvania utility.

FREE RIDER: A program participant who would have implemented the program measure or practice in the absence of the program.

GROSS SAVINGS: The change in energy consumption and/or demand that results directly from program-related actions taken by participants in an efficiency program, regardless of why they participated.

IMPACT EVALUATION: Used to measure the program-specific induced changes in energy and/or demand usage (such as kWh/year, kW, and therms) and/or behavior attributed to energy efficiency and DR programs.

IMPLEMENTATION CONSERVATION SERVICE PROVIDERS (ICSP): Contractors retained by an EDC to administer a specific EE&C program and generate ex ante savings values for efficiency measures.

INCENTIVES: Financial support (e.g., rebates, low-interest loans) to install energy efficiency measures. The incentives are solicited by the customer and based on the customer's billing history and/or customer-specific information.

INDEPENDENT VARIABLES: The factors that affect the energy and demand used in a building but cannot be controlled (e.g., weather, occupancy).

INTERNATIONAL PERFORMANCE MEASUREMENT AND VERIFICATION PROTOCOL (IPMVP): Defines standard terms and suggests best practices for quantifying the results of energy efficiency investments and increasing investment in energy and water efficiency, demand management, and renewable energy projects.

LOAD MANAGEMENT: Steps taken to reduce power demand at peak load times or to shift some of it to off-peak times. Load management may coincide with peak hours, peak days, or peak seasons. Load management may be pursued by persuading consumers to modify behavior or by using equipment that regulates some electric consumption. This may lead to complete elimination of electric use during the period of interest (*load shedding*) and/or to an increase in electric demand in the off-peak hours as a result of shifting electric use to that period (*load shifting*).

LOAD SHAPES: Representations such as graphs, tables, and databases that describe energy consumption rates as a function of another variable, such as time or outdoor air temperature.

MARKET EFFECT EVALUATION: The evaluation of the change in the structure/functioning of a market or the behavior of participants in a market that results from one or more program efforts. Typically, the resultant market or behavior change leads to an increase in the adoption of energy efficient products, services, or practices.

MARKET TRANSFORMATION: A reduction in market barriers resulting from a market intervention, as evidenced by a set of market effects, that lasts after the intervention has been withdrawn, reduced, or changed.

MEASURE: An installed piece of equipment or system, or modification of equipment, systems, or operations on end-use customer facilities that reduces the total amount of electrical or gas energy and capacity that would otherwise have been needed to deliver an equivalent or improved level of end-use service.

MEASUREMENT: A procedure for assigning a number to an observed object or event.

MEASUREMENT AND VERIFICATION (M&V): Activities to determine savings for individual measures and projects. This differs from evaluation, which is intended to quantify program impacts.

METERING: The use of instrumentation to measure and record physical parameters for energy-use equipment. In the context of energy efficiency evaluations, the purpose of metering is to accurately collect the data required to estimate the savings attributable to the implementation of energy efficiency measures.

MONITORING: Recording of parameters – such as hours of operation, flows, and temperatures – used in the calculation of the estimated energy savings for specific end uses through metering.

NET PRESENT VALUE (NPV): The value of a stream of cash flows converted to a single sum in a specific year, usually the first year of the analysis. It can also be thought of as the equivalent worth of all cash flows relative to a base point called the present.

NET SAVINGS: The total change in load that is attributable to an energy efficiency program. This change in load may include, implicitly or explicitly, the effects of free-drivers, free riders, energy efficiency standards, changes in the level of energy service, participant and non-participant spillover, and other causes of changes in energy consumption or demand.

NET-TO-GROSS RATIO (NTGR): A factor representing net program savings divided by gross program savings that is applied to gross program impacts to convert them into net program load impacts.

NON-PARTICIPANT: Any consumer who was eligible but did not participate in an efficiency program in a given program year. Each evaluation plan should provide a definition of a *non-participant* as it applies to a specific evaluation.

NON-RESPONSE BIAS: The effect of a set of respondents refusing or choosing not to participate in research; typically, larger for self-administered or mailed surveys.

PARTIAL FREE RIDER: A program participant who would have implemented, to some degree, the program measure or practice in the absence of the program (For example: a participant who may have purchased an ENERGY STAR appliance in the absence of the program, but because of the program bought an appliance that was more efficient).

PARTICIPANT: A consumer who received a service offered through an efficiency program, in a given program year. The term *service* is used in this definition to suggest that the service can be a wide variety of services, including financial rebates, technical assistance, product installations, training, energy efficiency information, or other services, items, or conditions. Each evaluation plan should define *participant* as it applies to the specific evaluation.

PEAK DEMAND: The maximum level of metered demand during a specified period, such as a billing month or a peak demand period.

PEAK DEMAND SAVINGS: The average energy savings during a system's peak demand period.[108]

PHASE II: EE&C programs implemented by the seven EDCs in Pennsylvania subject to the requirements of Act 129 during the program years ending on May 31 in 2014, 2015, and 2016.

PHASE III: EE&C programs implemented by the seven EDCs in Pennsylvania subject to the requirements of Act 129 during the program years ending on May 31, 2017–2021.

PHASE IV: EE&C programs implemented by the seven EDCs in Pennsylvania subject to the requirements of Act 129 during the program years ending on May 31, 2022–2026.

[108] Stern, Frank and Justin Spencer. 2017. Chapter 10: Peak Demand and Time-Differentiated Energy Savings Cross-Cutting Protocol. Uniform Methods Protocol. Weblink

PHASE V: EE&C programs implemented by the four EDCs in Pennsylvania subject to the requirements of Act 129 during the program years ending on May 31, 2027–2031.

PJM: PJM Interconnection, LLC, is a regional transmission organization (RTO) that coordinates the movement of wholesale electricity in all or parts of 13 states and the District of Columbia.

PORTFOLIO: Either (a) a collection of similar programs addressing the same market (e.g., a portfolio of residential programs), technology (e.g., motor efficiency programs), or mechanisms (e.g., loan programs), or (b) the set of all programs conducted by one organization, such as a utility (and which could include programs that cover multiple markets, technologies, etc.).

PRECISION: The indication of the closeness of agreement among repeated measurements of the same physical quantity.

PROCESS EVALUATION: A systematic assessment of an energy efficiency program for the purposes of documenting program operations at the time of the examination and identifying and recommending improvements to increase the program's efficiency or effectiveness for acquiring energy resources while maintaining high levels of participant satisfaction.

PROGRAM: A group of projects, with similar characteristics and installed in similar applications. Examples could include a utility program to install energy efficient lighting in commercial buildings, a developer's program to build a subdivision of homes that have photovoltaic systems, or a state residential energy efficiency code program.

PROGRAM YEAR: For Act 129, begins on June 1 and ends on May 31 of the following calendar year; impacts are reported annually. Program years are mapped to the PJM delivery year, not to the calendar year.

PROJECT: An activity or course of action involving one or multiple energy efficiency measures, at a single facility or site.

REGRESSION ANALYSIS: Analysis of the relationship between a dependent variable (response variable) and specified independent variables (explanatory variables). The mathematical model of their relationship is the *regression equation*.

RELIABILITY: Refers to the likelihood that the observations can be replicated.

REPORTING PERIOD: The time following implementation of an energy efficiency activity during which savings are to be determined.

RETROFIT ISOLATION: The savings measurement approach defined in IPMVP Options A and B, and ASHRAE Guideline 14, that determines energy or demand savings through the use of meters to isolate the energy flows for the system(s) under consideration.

RIGOR: The level of expected confidence and precision. Greater levels of rigor increase confidence that the results of the evaluation are both accurate and precise.

SPILLOVER: Reductions in energy consumption and/or demand caused by the presence of the energy efficiency program, beyond the program-related gross savings of the participants. There can be participant and/or non-participant spillover.

STIPULATED VALUES: An energy savings estimate per unit, or a parameter within the algorithm designed to estimate energy impacts that are meant to characterize the average or expected value within the population.

STATEWIDE EVALUATOR (SWE): The independent consultant under contract to the PUC to complete a comprehensive evaluation of the Phase V EE&C programs implemented by the four EDCs in Pennsylvania subject to the requirements of Act 129.

STATEWIDE EVALUATION TEAM (SWE TEAM): The team, led by Demand Side Analytics, that is conducting the evaluations of the Phase V Act 129 programs. Team members are Demand Side Analytics LLC, NMR Group Inc., Brightline Group, and Fulcrum Energy Solutions.

TECHNICAL REFERENCE MANUAL (TRM): A resource document that includes information used in program planning and reporting of energy efficiency programs. It can include savings values for measures, engineering algorithms to calculate savings, impact factors to be applied to calculated savings (e.g., NTG ratio values), source documentation, specified assumptions, and other relevant material to support the calculation of measure and program savings. It can also include the application of such values and algorithms in appropriate applications.

TIME-OF-USE (TOU): Electricity prices that vary depending on the time periods in which the energy is consumed. In a time-of-use rate structure, higher prices are charged during utility peak-load times. Such rates can provide an incentive for consumers to curb power use during peak times.

TECHNICAL UTILITY SERVICES (TUS): The bureau within the PUC that serves as the principal technical advisory staffing resource regarding fixed and transportation utility regulatory matters, as well as an adviser to the PUC on technical issues for electric, natural gas, water, wastewater, and telecommunications utilities.

UNCERTAINTY: The range or interval of doubt surrounding a measured or calculated value within which the true value is expected to fall within some degree of confidence.

UNIFORM METHODS PROJECT (UMP): Project of the U.S. Department of Energy to develop methods for determining energy efficiency for specific measures through collaboration with energy efficiency program administrators, stakeholders, and EM&V consultants – including the firms that perform up to 70% of the energy efficiency evaluations in the United States. The goal is to strengthen the credibility of energy efficiency programs by improving EM&V, increasing the consistency and transparency of how energy savings are determined.

VALUE OF INFORMATION (VOI): A balance between the level of detail (rigor) and the level of effort required (cost) in an impact evaluation.

Appendix B Common Approach for Measuring Net Savings for Appliance Retirement Programs

ARPs typically offer some mix of incentives and free pickup for the removal of old-but-operable refrigerators, freezers, dehumidifiers, or room air-conditioners. These programs are designed to encourage the consumer to do the following:

- Discontinue the use of secondary or inefficient appliances
- Relinquish appliances previously used as primary units when they are replaced (rather than keeping the old appliance as a secondary unit)
- Prevent the continued use of old appliances in another household through a direct transfer (giving it away or selling it) or indirect transfer (resale on the used appliance market)

Because the program theory and logic for appliance retirement differs significantly from standard downstream incentive programs (which typically offer rebates for the purchase of efficient products), the approach to estimating free ridership is also significantly different. Consistent with the Pennsylvania TRM, which relies on the U.S. Department of Energy UMP as the default inputs for estimating gross savings, the SWE Team recommends that the Pennsylvania EDCs also follow the UMP guidelines for estimating program net savings.[109] It is important to note that appliance replacement (with early retirement) programs are extensions of ARPs. Many of the principles described in this appendix will also apply to appliance replacement programs. For EDCs offering appliance replacement programs, their evaluation plans should draw upon this Appendix in proposing their approach to assessing the net impacts of the programs.

In the following sections, we present the UMP approach, adding in clarifying explanations/diagrams where applicable. Note that this is based on the current version of the UMP that no longer includes an induced replacement adjustment as part of the net savings calculations.

B.1 General Free Ridership Approach

The nature of the ARP requires a unique approach to estimating free ridership, and ultimately, net savings. Free ridership is based on the participants anticipated plans had the program not been available – a free rider is classified as one who would have removed the unit from service irrespective of the program. Net savings for the ARP is therefore based on the participants' anticipated continued operation of the appliance either as a primary or a secondary unit, within their home or transferred to another home (either directly or indirectly).

---

109 See Keeling, Josh and Doug Bruchs. 2017. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures, Chapter 7: Refrigerator Recycling Evaluation Protocols, National Renewable Energy Laboratory, November 2017. Weblink

The general approach to estimating net savings for an ARP is to segment participants into three different groups or scenarios of what would have happened to a program-recycled unit in the absence of the program:

1. The household would have kept the unit or given it directly to a close acquaintance.
2. The unit would have been transferred directly or indirectly to a customer (other than a close acquaintance) for continued use.
3. The unit would have been discarded by a method that would lead to its permanent removal from service.

To categorize a participant into one of the three scenarios, evaluators should ask participants what they would have done with the appliance in the absence of the program. Table 43 provides common response options, scenario assignment, and free ridership status.

Table 43: Free Rider Scheme

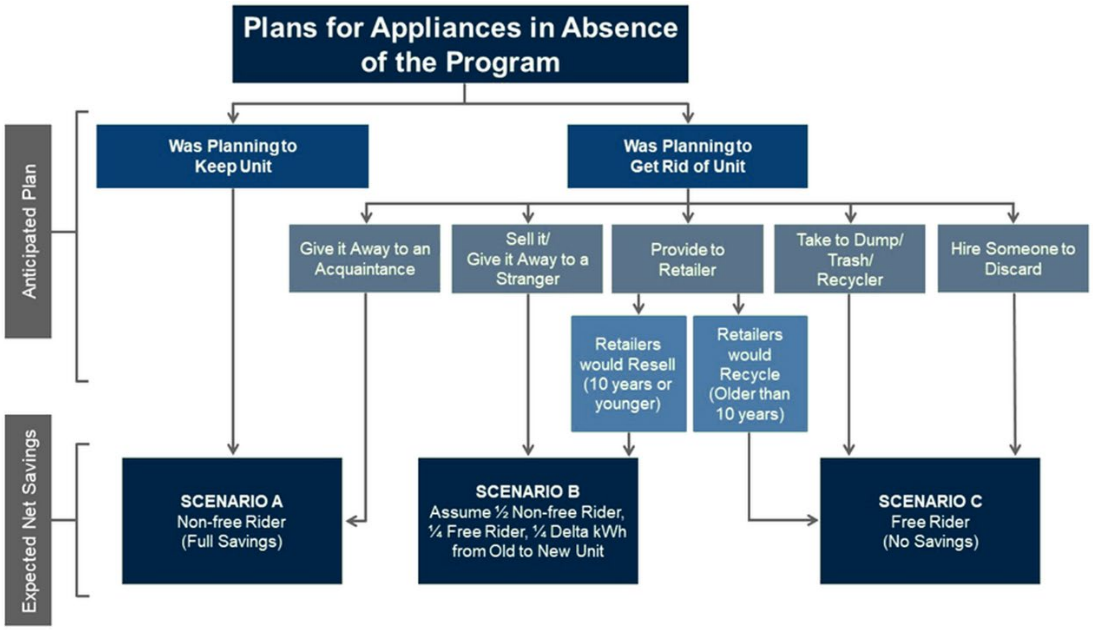
Self-Reported Alternatives to the Program	Scenario	Free Ridership Status
Kept by the household	Scenario A	Not a Free Rider
Given away for free to an acquaintance	Scenario A	Not a Free Rider
Sold; given to charity	Scenario B	See Algorithm in Figure 24
Provided to retailer & ten years old or newer[1]	Scenario B	See Algorithm in Figure 24
Provided to retailer & older than ten years[1]	Scenario C	Free Rider
Hauled to landfill or recycling center; hired someone to discard	Scenario C	Free Rider

[1]The ten-year age cutoff for resale value was derived from the following study: Navigant Consulting, January 22, 2013: Energy Efficiency/Demand Response Plan: Plan Year 4 Evaluation Report: Residential Fridge and Freezer Recycle Rewards Program; Prepared for Commonwealth Edison Company

The free ridership algorithm is depicted visually in Figure 24. The algorithm was developed based on the UMP guidance.[110] The algorithm assigns respondents who planned to keep their units or give them to acquaintances as non-free riders; these respondents receive full savings (Scenario A). Free riders include anyone who planned to dispose, recycle, or discard the unit, or to provide an older unit to a retailer (Scenario C). For those participants who planned to transfer the unit to another user by selling it, giving it away to charity or stranger, or providing younger units to retailers (Scenario B), the SWE recommends, based on the UMP, assuming that one-half of the units are not-free riders (and received full savings) and one-quarter of the units were free riders. The remaining quarter of transfers should be assigned the difference in savings between the verified gross energy savings (old unit) and the weighted average consumption of newly manufactured units (kWh_ee).

[110] The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures, Chapter 7: Refrigerator Recycling Evaluation Protocols, National Renewable Energy Laboratory, November 2017. Weblink. See also NMR Group. 2018. Appliance Recycling Report. Weblink

Figure 24: Free Ridership Algorithm[1]



[1] Algorithm figure originally reported in NMR Group. 2018. Appliance Recycling Report. Visual depiction of UMP recommendations provided by Scott Dimetrosky

B.2 Estimating Net Savings

Net savings should be assigned individually to each respondent based on the responses to a participant survey and categorization to the scenarios as outlined above. The net savings should be averaged across all respondents to calculate program-level net savings. Table 44 demonstrates the proportions of a sample population that are classified into each of the potential scenarios and the resulting weighted net savings.

Table 44: Net Savings Example for a Sample Population*

Scenario	Free ridership Status	Population (%)	UEC (kWh) w/out Program	UEC (kWh) w/ Program	kWh Savings
Scenario A (kept unit)	Not a free rider	50%	1,000	0	1,000
Scenario B (sold, donated, provided to retailer)	Assumed non-free rider (1/2 of Scenario B)	15%	1,000	0	1,000
	Assumed free rider (1/4 of Scenario B)	8%	0	0	0
	Delta kWh from old to new unit (1/4 of Scenario B)	8%	1,000	500	500
Scenario C	Free rider	20%	0	0	0
Avg Net Savings (kWh)			688		

* The percent values presented in this table are just examples; actual research should be conducted to determine the percentage of units that fall into each of these categories. The UEC values presented in the table are also for example only. EDCs should use the 2026 PA TRM to determine the UEC of retired units.

B.3 Data Sources

A random sample survey of program participants should be the primary source of data collected for estimating NTG for the appliance recycling program. Per the UMP, a secondary source of supporting data may come from a non-participant sample survey. Non-participants do not have the same perceived response bias as participants and can help offset some of this potential bias in estimating the true proportion of the population that would have recycled their unit in the absence of the program. To maintain consistency with the UMP, we recommend averaging the results of the non-participant survey with those of the participant survey. The use of a non-participant survey is recommended but not required given budget and time considerations.

Appendix C Common Approach for Measuring Free Riders for Downstream Programs

C.1 Introduction

The PA PUC Implementation Order specifies that the NTG ratio for Phase V of Act 129 is to be treated in the same way as previous Phases. Specifically, for compliance purposes, the NTG ratios for Phase V programs continue to be set at 1.0 – basing compliance with energy and demand reduction targets on gross verified savings. However, the PUC order also states that the EDCs should continue to use net verified savings to inform program design and implementation.

There are two reasons to consider having a uniform NTG approach for the EDCs. One is that if NTG measurement for a program is consistent across time, comparisons of the NTG metric across time will be reliable and comparisons are therefore valid. If the NTG metric is measured the same way every year or every quarter, program staff can use the NTG metric to inform their thinking because it provides a consistent metric over time. Of course, programs often change across years: measures may be added or taken away, and rebate amounts or technical services may vary. Consistent measurement of NTG is even more valuable in these situations because it permits better understanding of how the changes affect NTG.

The second reason to consider having a uniform NTG approach for the EDCs is the value that can be obtained from comparisons across utilities. Just as programs change year to year, the programs offered by the EDCs vary from each other. When there are different metrics, no one can discern whether different NTG values are due to program differences, external differences, or differences in the metric. By using a consistent metric, we can at least rule out the latter.

The variability in the types of services/measures offered by the programs, the different delivery strategies, and the variability of the customer projects themselves makes it necessary to tailor the attribution assessment appropriately. The need for comparability of results between years and between EDCs, however, requires a consistent overall approach to assess attribution. The challenge is in allowing flexibility/customization in application yet still maintaining a consistent approach.

C.2 Sources for Free Ridership and Spillover Protocols

Care must be taken when developing the questions used to measure free ridership. The SWE considers the research approaches detailed in the UMP[111] as well as those used in Massachusetts[112] and those

---

[111] Violette, Daniel and Pamela Rathbun, "Estimating Net Savings: Common Practices," in The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. Prepared for the National Renewable Energy Laboratory, October 2017. Weblink
[112] Tetra Tech; KEMA; NMR Group, Inc. 2011. Cross-Cutting (C&I) Free ridership and Spillover Methodology Study Final Report. Massachusetts Program Administrators. Weblink

developed by the Energy Trust of Oregon[113] to constitute some of the best practices for free ridership and spillover estimation.

The Framework provides the following general guidance as a good starting place for assessing free ridership and spillover. Furthermore, the SWE recommends standardization – at a minimum within the EDCs' measurement activities and ideally across all EDCs – for provision of consistency in explaining program effects. Among several free ridership methods mentioned, the SWE recommends an approach similar to that chosen by the Energy Trust, which uses a concise battery of questions to assess intention and program influence, which is the focus of the rest of this appendix. [114]

The Framework also defines participant and non-participant spillover and recommends the consideration of trade ally surveys and reports for assessing the non-participant portion of a program's spillover impact.

C.3 Sampling

The sampling approach for estimating free riders should use confidence and precision levels at least equivalent to the approach for gross savings being estimated for a specific program. The SWE further recommends sampling and reporting free ridership and spillover by stratifying for high-impact end-uses in much the same way as for gross savings estimates whenever possible (see Section 3.4.1.4). EDCs are encouraged to use higher confidence and precision levels, and to conduct the sampling at the measure level when more detailed information is needed for program assessment.

C.4 Recommended Standard Free Ridership Protocol

The following discussion presents a standard, yet flexible, approach to assessing free ridership for the EDCs to use during Phase V. This method applies to downstream programs, typically using some incentive or direct installation.[115] Research Into Action and Energy Trust of Oregon developed this

---

NMR Group, Inc. and Tetra Tech (2011). Cross-Cutting Net to Gross Methodology Study for Residential Programs –Suggested Approaches. Weblink
TetraTech 2017. Net-to-Gross Methodology Research. Weblink
NMR Group and Tetra Tech. 2020. Consistent Methodology for Self-Reported Residential Net-to-Gross Measurement. Weblink
NMR Group and Tetra Tech. 2022. Residential Self-report Net-to-Gross Methodology Update and Sensitivity Analysis. Weblink
NMR Group and Tetra Tech. 2026. Commercial & Industrial Self-report Net-to-Gross Survey and Algorithm Update. Weblink
NMR Group and Tetra Tech. 2026. Residential Self-report NTG Methodology. Weblink

[113] Energy Trust Free Ridership Methodology. Weblink

[114] Ibid.

115 When self-report questions are used for upstream and mid-stream programs those questions should use the same structure described herein. However, self-report methods are typically insufficient and additional data sources should be used but are not prescribed at this time.

approach for telephone and on-site assessment of NTG (by project and by measure) across residential, commercial, industrial, and government sectors, including the following:

- Rebates and grants for energy efficiency improvements
- Rebates and grants for renewable energy sources
- Technical assistance
- Education and outreach

The assessment battery is brief to avoid survey burden yet seeks to reduce self-report biases by including two components of free ridership: (1) intention to carry out the energy efficient project without program funds and (2) influence of the program in the decision to carry out the energy efficient project. When scored, each component has a value ranging from zero to 50, and a combined total free ridership score that ranges from zero to 100. These components are potentially subject to different and opposing biases. As a result, the intention component typically indicates higher free ridership than the influence component. Therefore, combining those decreases the biases.

For Phase V, because of the Commission's emphasis on collaboration with other program administrators, we recommend adding to the assessment of free ridership the impact of other programs on both intention and influence by (1) determining if a project included outside funding and (2) if so, asking respondents to attribute the impacts of the EDC program in isolation and the impacts of the combined EDC program and outside funding. This will provide the EDCs and stakeholders with additional information regarding the effectiveness of EDC programs as well as the EDC efforts to collaborate with outside funding.

In the following subsections, we describe a Common Method for a standard retrofit incentive program, including both the question battery and scoring. We describe how the Common Method can be adapted for different types or variations of program or measure types (e.g., EDC direct install and custom programs). We finally address several questions and concerns that EDCs and their ECs raised in response to earlier versions of this memo.

C.4.1 Intention

Intention is assessed through a few brief questions used to determine how the upgrade or equipment replacement likely would have differed if the respondent had not received the program assistance. The initial question asks the respondent to identify, of a limited set of options, the option that best describes what most likely would have occurred without the program assistance. Note that program assistance often includes more than just the incentive or rebate – it may also include audits, technical assistance, and the like.

To account for the impact of outside programs, we recommend including a question before the *intention* battery to determine if the respondent received non-EDC program assistance. If such assistance was received, respondents should be asked the intention battery of questions both for the EDC program in isolation and for the combined support from the EDC program and outside funding.

The response options offered (typically four or five, and preferably no more than six) capture the following general outcomes:

- Would have canceled or postponed the project, upgrade, purchase, etc., beyond the current program cycle (typically at least one year).
- Would have done something that would have produced savings, but not as much as those achieved through the upgrade or equipment replacement as implemented.
- Would have done the upgrade or equipment replacement as implemented.
- Don't know.

The first outcome (canceled or postponed beyond the program cycle) indicates zero free ridership and thus results in a score of 0. The second option indicates some free ridership, but not total free ridership (a score ranging from 12.5 to 37.5 for the intention component). The level of free ridership depends on two factors: (1) the level of savings that the respondent would have achieved without the program's assistance, and (2) in the case of non-residential programs, whether the respondent's business or organization would have paid the entire cost of the equipment replacement or upgrade without the program assistance. The third outcome (done project as implemented) indicates total free ridership (a score of 50 for the intention component).

In previous implementations of this approach, "don't know" responses to this question were assigned the midpoint score of 25 for the intention component. Alternative treatments that have been proposed for "don't know" responses are to assign the mean of non-missing responses or to exclude the case and replace it with another. Both those treatments may be problematic, as they assume that "don't know" responders are otherwise similar to the rest of the sample, when there may be reasons for the "don't know" response that make them dissimilar. Generally, imputing the mean for missing responses is not considered best practice.[116]

We recognize that imputing the midpoint may be considered arbitrary (but see Section below on treatment of "don't know" responses). Moreover, our experience is that "don't know" responses are infrequent, and so the way in which they are handled likely will not have a great impact on the resulting free ridership estimates. Evaluators may implement alternative approaches to handling "don't know" responses in addition to assigning the midpoint and report both results. As an alternative approach, we recommend using linear regression to predict the intention score from each respondent's influence score.

As discussed below, the assessment of the above factors will depend somewhat on the nature of the program, but the overall approach is guided by several considerations:

- The instrument should be as brief as possible to avoid survey burden.

[116] Enders, C.K. *Applied Missing Data Analysis*, New York: The Guilford Press, 2010.

- Challenging a respondent's consistency can make the respondent feel defensive and may not produce more accurate data; therefore, the instrument should avoid overt consistency checks.
- The instrument should recognize the limits of reporting a counterfactual, particularly in assessing cases in which respondents that report they would have saved some, but less, energy without the program.

Any tailoring of the approach should take the above considerations into account.

The following subsections describe, in turn, how intention typically has been assessed with the Common Method in non-residential and residential programs and how it can be further tailored if needed.

C.4.2 Assessment of Intention in Non-Residential Programs

In this section, we describe how the Common Method typically is applied and scored in standard, non-residential incentive programs. We also discuss tailoring or modification of the Common Method.

General Application of Intention Assessment in Non-Residential Programs

Before the non-residential intention battery, we recommend including a question that determines if the project received any non-EDC assistance:

- In addition to the support from [EDC PROGRAM], did you receive financial support or technical support from a state or federal program such as [INSERT APPLICABLE PROGRAMS, SUCH AS "Reducing Industrial Sector Emissions in Pennsylvania (RISE PA)"]

The non-residential battery first asks about the EDC program in isolation and begins with the following question:

- Which of the following is most likely what would have happened if you had not received [the EDC program assistance]? [IF THE RESPONDENT INDICATED ADDITIONAL, OUTSIDE ASSISTANCE: "PLEASE CONSIDER ONLY THE EDC PROGRAM ASSISTANCE."]

The battery has included the following options in multiple evaluations of a wide range of non-residential programs:

- Canceled or postponed the project at least one year
- Reduced the size, scope, or efficiency of the project
- Done the exact same project
- Don't know

Respondents that select the second option are asked the following:

- By how much would you have reduced the size, scope, or efficiency? Would you say…

  a. a small amount,
  b. a moderate amount, or

c. a large amount

If the respondent receives outside assistance, we recommend repeating the intention battery, and beginning with the following type of question and using the same response options: [117]

- Which of the following is most likely what would have happened if you had not received both [the EDC program assistance] and the additional support you mentioned previously?

Note that the intent is not to separately assess reduction in size, scope, and efficiency – it is simply to assess whether, in the respondent's opinion, in absence of the program the project would have been reduced in size, scope, or efficiency by a small, moderate, or large amount. Under the above assumption that a precise estimate of counterfactual savings is not likely to be achievable, this approach makes no effort to establish such an estimate. Instead, the approach simply attempts to obtain the respondent's best general estimate of the counterfactual. The SWE notes that a large reduction in a given project's size would not necessarily have the same energy impact as a small, moderate, or large reduction in the project's scope or the efficiency level of the equipment used. However, the purpose is to balance the desire to obtain some estimate of savings reduction with the desire to avoid response burden and reduce the risk of false precision.

Nevertheless, evaluators may propose alternative response options. The SWE requests that those evaluators provide their rationale for such alternatives.

Respondents who report they would have done exactly the same project without the program's assistance are asked the following:

- Would your business have paid the entire cost of the upgrade?

Respondents who report they would have done exactly the same project without the EDC program's assistance and the outside assistance are asked the following:

- Would your business have paid the entire cost of the upgrade, without the assistance from the EDC program and the additional support you mentioned previously?

This question is used to help mitigate a bias to overstate the likelihood that the respondent would have done the same project without program assistance.[118] Respondents get the highest free rider score only if they report that they would have done the same project without program assistance and that their business would have paid the entire cost. Otherwise, a lower free rider score is assigned, as shown below.

It is important to note that the above question is not a consistency check. That is, respondents who report they would have done the same project without program assistance but do not confirm that their business would have paid the entire cost are not confronted with the apparent inconsistency and

[117] Respondents who did **not** report receiving outside assistance would not be asked a second intention battery.
118 See Section C.6.1, Controlling for Socially Acceptable Response Bias, for a more complete discussion of this potential bias.

asked to resolve it. Nor does the method assume that the second response is the correct one. Instead, the method assumes that neither response provides the full picture, and that further questioning could not reliably provide the complete picture. The method thus assigns a free rider value that is intermediate to both. That is, it assumes that the best estimate is that the project would have produced some savings but not as much as were actually produced through the program.

Scoring of Intention Assessment in Non-residential Programs

An intention free ridership score of 0 to 50 is assigned as follows:

- A project that would have been canceled or postponed beyond the program cycle is assigned an intention score of 0.
- A project that would have been done exactly as it actually was done, with the cost born entirely by the respondent's business or organization, is assigned an intention score of 50.
- A project that would have resulted in fewer savings than the project actually done is assigned an intermediate score based on the responses to the applicable follow-up question(s).

Interviewers (or web surveys) should make reasonable attempts to get a response to the questions. If respondents cannot select an option, "don't know" responses are assigned a score that represents the midpoint of the range of possible values for that question (as illustrated below).[119]

Table 45 summarizes the possible response combinations to the questions described above and the intention score assigned to each unique combination. The intention scoring is identical for respondents who indicate receiving outside program support.

Table 45: General Free Ridership Intention Component Scoring

Question	Response	Intention Score
1. Which of the following is most likely what would have happened if you had not received [the program assistance]?	Postponed / cancelled	0
	Reduced size, scope, efficiency	Based on response to Q2
	No change	Based on response to Q3
	Don't know	25*,**
2. By how much would you have reduced the size, scope, or efficiency?	Small amount	37.5
	Moderate amount	25
	Large amount	12.5
	Don't know	25*
3. Would your business have paid the entire cost of the upgrade?	Yes	50
	Don't know	37.5*
	No	25**

[119] Section C.6.3, *Treatment of "Don't Know" Responses*, discusses the rationale for this treatment of "don't know" responses rather than alternatives, such as assigning a mean value. In fact, "don't know" responses are infrequent.

* Represents the midpoint of possible values for this question.

** Infrequent response.

Tailoring of Intention Assessment in Non-Residential Programs

The above approach has been used to assess intention with a range of retrofit incentive programs. Evaluators may propose other modifications as needed, but such modifications should be informed by the general principles described above, of keeping the instrument brief, recognizing the limits of counterfactual questioning, and avoiding consistency checks.

Tailoring of Question Wording

The specific wording of the questions and the response options provided should be tailored to the specific program, measure type, or sample group. As indicated above, the general form of the initial intention question is "Which of the following is most likely what would have happened if you had not received [the program assistance]?" Therefore, it is important to identify the primary type or types of program assistance that are considered important in reducing the key barriers to carrying out the targeted behavior (e.g., an upgrade to more energy efficient equipment). In other words, it is important to clearly indicate what participating in the program meant and what program they were participating in.

Example: A program operated through a State agency helped businesses obtain contracts with an Energy Services Company (ESCO) to finance efficiency upgrades. In this case, the intention question was as follows:

> *What do you think your organization most likely would have done if the [Name of Office] had not helped you obtain the contract with an ESCO like …?*

As noted above, the influence question should include the range of program elements or services. Evaluators should be careful not to ask about services that a particular program does not provide. For example, it would be confusing to ask how influential the rebate was if there was no rebate attributable to the program/measure. Logic models, program theory, and staff interviews typically inform the list of program elements to ask about.

Tailoring of Response Options

As noted above, one area in particular where modification may be proposed is in the specification of equipment replacement or upgrade alternatives to identify differing levels of counterfactual energy savings (i.e., in place of asking whether the respondent would have done something that reduced energy by a small, moderate, or large amount). In such cases, the counterfactual options should reflect the range of activities that likely would have occurred absent program assistance, with points assigned to reflect the amount of energy savings each would provide.

For example, the following alternatives could be specified for a lighting program that incents LED fixtures:

1. Put off replacing the [X type of] fixtures with LED fixtures for at least one year or cancelled it altogether.

2. Kept some of the existing fixtures and replaced some fixtures with LED fixtures.
3. Installed different lighting fixtures. If so, what kind? ____________________
4. Installed the same number and type of LED fixtures anyway.
5. Done something else. If so, what? ____________________
6. Don't Know or no answer.

Follow-up questions are needed for some responses. In this case, for respondents who report they would have installed fewer lights, a follow-up question is needed to assess the savings reduction – specifically, what percentage of fixtures would they have replaced with LED fixtures? For respondents who said they would install the same number, a follow-up question should be used to verify that the respondent would have paid the entire cost without program support.

Other Tailoring or Modifications

Examples of additional types of modifications include the following:

- Preceding the initial counterfactual question with one asking whether the respondent had already carried out the equipment replacement or upgrade before applying for the incentive.
  - Evaluators may include such a question but should still ask the counterfactual question as described above.
- Specifying the value of each respondent's incentive in the initial counterfactual question.
  - This is acceptable, but evaluators should keep in mind that the incentive often is not the only program assistance received and other program assistance may also have had a role in driving the project. So, for example, the question may refer to "the incentive of $X and other assistance, such as identification of savings opportunities."

We provide further discussion of tailoring the general free ridership approach for programs other than standard retrofit type programs below.

C.4.3 Assessment of Intention in Residential Programs

The assessment of intention for residential programs is similar to that for non-residential programs. However, the response option "reduced the size, scope, or efficiency of the project" is not likely to be as meaningful to a residential respondent as to a non-residential one, nor is a residential respondent expected to be able to estimate whether the reduction would be small, moderate, or large. Evaluators, rather, should attempt to provide a list of meaningful counterfactual options.

As with the non-residential battery, we recommend including a question that determines if the project received any non-EDC assistance before asking the intention battery of questions, and if the respondent received outside assistance, we recommend repeating the intention battery for the combined effects of EDC assistance and non-EDC assistance on intention.

Table 46 shows examples of counterfactual response options used with three types of residential measures: appliances, air or duct sealing or insulation, and windows. As this shows, the goal is to cover the range of likely alternatives to carrying out the incented upgrade, with intention scores that reflect the degree of free ridership. Reporting an alternative that likely would have produced no energy savings results in a score of 0; reporting something that likely would have produced some energy savings, but lower savings than the incented upgrade or purchase results in an intermediate score of 0.25; and reporting the same outcome as the incented upgrade or purchase results in a score of 0.5.

Table 46: Example Counterfactual Response Options for Various Residential Measure Types

Program	Counterfactual Responses	Intention Score
Appliance	Cancel/postpone purchase	0
	Repair old appliance	0
	Buy used appliance	0
	Purchase less expensive appliance	0.25
	Purchase less energy efficient appliance	0.25
	Purchase same appliance without the rebate	0.5
	Don't know	0.25
Air/Duct Sealing, Insulation	Cancel/postpone	0
	Do by self (if program incents only contractor-installation)	0.25
	Reduce amount of sealing/insulation	0.25
	Have the same level of sealing/insulation done without the rebate	0.5
	Don't know	0.25
Windows	Cancel/postpone purchase	0
	Replace fewer windows	0.25
	Purchase less expensive windows	0.25
	Purchase less energy efficient windows	0.25
	Do same window replacement without the rebate	0.5
	Don't know	0.25

A difference from the non-residential instrument is that respondents who report they would have done the same thing without the incentive are not then asked whether they would have paid the cost of the upgrade. A question that may seem perfectly reasonable in the context of a decision about allocating a business's resources may not seem reasonable in the context of personal decisions. Instead, the "would have done the same thing" response may include the words "without the rebate [or incentive]."

Issues relating to tailoring the intention component are the same as for non-residential assessments.

C.4.4 Influence (Non-Residential and Residential)

Assessing program influence is the same for non-residential and residential programs.

Program influence may be assessed by asking the respondent how much influence – from 1 (no influence) to 5 (great influence) – various program elements had on the decision to do the project the way it was done.

The number of elements included will vary depending on program design. Logic models, program theory, and staff interviews typically inform the list. The more typical elements programs use to influence customer decision making include information; incentives or rebates; interaction with program staff (technical assistance); interaction with program proxies, such as members of a trade ally network; building audits or assessments; and financing.

If the respondent received outside assistance, we recommend including at least one element for the outside assistance.

The program's influence score is equal to the maximum influence rating for any program element rather than, say, the mean influence rating. The rationale is that if any given program element had a great influence on the respondent's decision, then the program itself had a great influence, even if other elements had less influence. If the respondent received outside assistance, the EDC program influence score is equal to the maximum influence rating for any EDC program element. A second program influence score is equal to the maximum influence rating for any program element (EDC or non-EDC).

Table 47: General Free Ridership Influence Component

Calculation of the Influence Score is demonstrated in the following example:							
Rate influence of program elements							
	Not at all influential				Extremely influential		
Incentive	1	2	3	4	5	DK	NA
Program staff	1	2	3	4	5	DK	NA
Audit/study	1	2	3	4	**5**	DK	NA
Marketing	1	2	3	4	5	DK	NA
Etc.	1	2	3	4	5	DK	NA

In this example the highest score (a 5 for the influence of the audit/study) is used to assign the influence component of the FR score. High program influence and FR have an inverse relationship – the greater the program influence, the lower the free ridership, as seen in Table 48.

Table 48: General Free Ridership Influence Component Scoring

Program Influence Rating	Influence Score
1 – not at all influential	50
2	37.5
3	25
4	12.5
5 – extremely influential	0
DK	25

C.4.5 Total Free Ridership Score

Total free ridership is the sum of the intention and influence components, resulting in a score ranging from 0 to 100. This score is multiplied by 0.01 to convert it into a proportion for application to gross savings values.

If a program includes participants who received outside assistance, the evaluation should include a second free ridership score that is derived from the supplemental intention and influence components that considered outside assistance among those participants. The purpose of this secondary free ridership score is meant to provide the EDCs and stakeholders with additional information regarding the effectiveness of EDC efforts to collaborate with outside funding.

C.5 Applying the Common Method to Other Program Types

Evaluators should be able to use the Common Method, described above, with most retrofit incentive programs. Evaluators may tailor the approach for use with programs that do not fit the general retrofit incentive mold.

In programs where the primary program approach is to provide assistance (e.g., rebate/incentive, technical assistance, direct install) to the program participant to reduce barriers to undertaking energy efficient upgrades or improvements, it typically should be sufficient to tailor question wording and response options while maintaining the overall approach. In such cases, the intention component may require more tailoring than the influence component.

In programs that must influence multiple actors to achieve the desired outcomes or carry out their influence through more complex forms of assistance, it may be necessary to tailor the method more extensively or to propose an alternative approach. Section C.6.1 discusses the process for proposing methods in the above cases.

The following examples show how the method has been applied for some programs that do not fit the standard retrofit incentive model. The purpose of these examples is not to show the only possible ways in which the Common Method may be modified to use with different program types but are here for illustrative purposes. EDCs and their evaluators should propose an approach that is consistent with the considerations outlined in Section C.4.1, above.

The first example illustrates a case for which the modification is relatively simple; the second example illustrates a more complex case requiring more extensive modification.

C.5.1 Direct Install Program

Direct install programs are different from most programs in that the program is offered directly to potential participants via program representatives. In applying the Common Method to a Direct Install program, the battery should verify whether the respondent was even considering the directly installed measure(s) prior to program contact. Where the respondent was not even considering the measures before being contacted by the program, the total free ridership score is set to 0 (i.e., both the intention and influence scores were 0). For respondents who were planning an upgrade, the method mirrors the general approach described above.

Assessment of program influence should be as described above but include potential program influences reflecting the unique elements of the Direct Install program. For example, in a case where the program included a building assessment along with Direct Install measures, the influence question should include "assessment results," along with "interactions with the assessor or contractor," and "the fact that the measure was free."

C.5.2 Financing an Energy Performance Contract

Some programs will require more extensive and ad hoc tailoring of the Common Method, such as when a program works with third-party entities to assist with project financing. In one example, a program helped building owners establish and implement energy performance contracts (EPCs) with program-administrator-approved energy service companies (ESCOs). Since the program administrator worked with both the building owner and the ESCO, neither alone could accurately describe what would have happened without the assistance. Therefore, for each sampled project, the evaluator should survey both the building owner and the ESCO.

The building owner instrument should include the standard intention question of what would have happened (postpone/cancel, smaller project, same upgrade) without program support and the standard influence question.[120] The evaluator should calculate building owner *intention* and *influence* following the standard approach, described above.

The instrument for ESCOs should ask the following:

- How likely they would have known about the client without the program's assistance.
- What likely would have happened without the program's assistance (same EPC, lower-savings EPC, no EPC).

The evaluator should calculate only ESCO intention, using the algorithm shown in Table 49.

---

[120] Examples of influencers include program information, interaction with program staff, the list of prequalified ESCOs, and program assistance in selecting an ESCO.

Table 49: Algorithm for ESCO Intention Score

Would Likely Have Known About Client	Counterfactual	Intention Score
Yes, likely would have known about client's needs without program assistance	Same EPC	50
	Lower-savings EPC	25
	No EPC	0
No, likely would not have known about client's needs without program assistance	N/A	0

To aid in determining how to combine the building owner and ESCO scores, the building owner instrument should ask the following:

- Whether they had ever worked with an ESCO before
- Whether they would have used an ESCO without program assistance

The evaluators in this example use the algorithm shown in Table 50 to calculate the intention component score based on responses by both the building owner and the ESCO. The algorithm assumes that the ESCO responses were not relevant if (1) the building owner was experienced with ESCOs and so could accurately predict what would have happened without the program assistance, and (2) the owner indicated that without program assistance they would have cancelled or postponed the project or would not have used an ESCO.

Table 50: Algorithm for Combining Building Owner and ESCO Intention Score

Would Have Used ESCO?	Bldg. Owner experienced with ESCO	ESCO responses considered?	Bldg. Owner Response to Intention Questions	ESCO Response to Intention Questions	Final intention score
No/DK	N/A	No[a]	Free Rider, Partial or Not Free Rider	N/A	Client score
Yes	Yes	No[b]			
Yes	No	Yes	Free Rider (would have done same project)	Free Rider	50
				Partial Free Rider	37.5
				Not Free Rider	25
			Partial Free Rider (would have done less efficient project)	Free Rider	25
				Partial Free Rider	25
				Not Free Rider	12.5
		No[c]	Not Free Rider (would have cancelled or postponed)	N/A	0

[a] Since the building owner would not have used an ESCO without program assistance, ESCO responses are not relevant.
[b] Since the building owner was experienced with ESCOs, it was assumed that they could accurately predict what would have happened without program assistance.

[c] Since the building owner indicated they would have cancelled or postponed the project without program assistance, the ESCO responses are not relevant.

In other cases, where there may be reason to question the building owner's ability to provide an accurate intention response, then the ESCO's response would also be considered and could be used to adjust the building owner's score.

C.6 Response to Questions and Concerns Raised About the Common Method

This section provides responses to questions and concerns about the Common Method raised in previous Phases of Act 129. We also provide additional information and clarification here in reference to specific questions or concerns raised.

C.6.1 Controlling for Socially Acceptable Response Bias

One concern is that respondents' self-reports are likely to be tainted by a bias toward reporting that they would have done the energy-saving project even without the program. This assumption has variously been ascribed to a social desirability bias (where energy conservation is the socially desirable response) or to an attribution bias (in which we tend to make internal attributions for good decisions or outcomes and external attributions for poor ones).

Above, we argued that the two components of free ridership that the battery assesses – intention to carry out the energy efficient project and influence of the program – are likely subject to different and opposing biases, which are at least partly canceled out by combining the components. While the intention component is subject to biases that would increase the estimate of free ridership, the influence component may be subject to biases that would decrease the estimate of free ridership. Specifically, rated influence may reflect satisfaction with the program such that participants who are satisfied with the program may report greater program influence. If so, a program with high participant satisfaction may appear to have lower free ridership on that basis.

Analysis of responses to the battery tend to support the above suppositions. In previous research, members of the SWE analyzed responses to the battery from 158 participants in non-residential retrofit and new construction programs and 1,252 participants in a range of residential programs (appliances, shell measures, home performance, and refrigerator recycling).[121] First, the two components positively correlated in both the non-residential and residential samples (.40 and .37, respectively), indicating shared measurement variance. However, the intention component yielded higher mean scores than did the influence component for both the non-residential (95% confidence interval: 16.8 ± 3.4 vs. 5.3 ± 1.5) and residential (95% confidence interval: 26.4 ± 1.3 vs. 10.5 ± 0.8) samples. If the shared variance

[121] The responses were collected in May through July of 2010, as part of the evaluation of roll-out of the Energy Trust Fast Method for collecting participant feedback. *Fast Feedback Program Rollout: Non-residential & Residential Program Portfolio*. Submitted to Energy Trust of Oregon by Research Into Action, Inc., December 31, 2010.

between the two components indicates they are both measuring free ridership, these findings are consistent with the idea that intention may over-estimate free ridership and influence may under-estimate it. Absent any compelling evidence that one of these components by itself yields a truer estimate of free ridership, it is safest to conclude that combining them provides the best assessment.

C.6.2 Intention Counterfactual Indicates Reduced Energy Savings

The Common Method provides three counterfactual options: (1) the upgrade would have been canceled or postponed at least one year; (2) the upgrade's size, scope, or efficiency would have been reduced; and (3) the same upgrade would have been done. Respondents who report a reduction in size, scope, or efficiency are then asked whether the reduction would be small, moderate, or large.

Three questions have been raised about the treatment of a reported reduction in size, scope, or efficiency:

- Does the method ask separately about the reduction in size, in scope, and in efficiency and, if so, how does it combine or weight the responses?
- Does the Common Method allow for asking about specific changes in size, scope, or efficiency? For example, in the case of a lighting project, could the instrument ask if the respondent would have installed different kinds of lights and, if so, what kind?
- If the Common Method allows for asking about specific changes in size, scope, or efficiency, how should the response be scored if the respondent does not provide enough information to determine a counterfactual difference in energy savings?

The underlying concern is whether the approach is capable of accurately capturing the difference in energy savings between the project-as-implemented and the counterfactual case where some energy savings would have been achieved.

As noted above, the intent is not to separately assess reduction in size, scope, and efficiency – it is simply to assess whether, in the respondent's opinion, in absence of the program the project would have been reduced in size, scope, or efficiency by a small, moderate, or large amount. Under the assumption that a precise estimate of counterfactual savings is not likely to be achievable, this approach makes no effort to establish such an estimate. Instead, the approach simply attempts to obtain the respondent's best general estimate of the counterfactual.

It is understood that a small, moderate, or large reduction in a given project's size would not necessarily have the same energy impact as a small, moderate, or large reduction in the project's scope or the efficiency level of the equipment used. The purpose is to balance the desire to obtain some estimate of savings reduction with the desire to avoid response burden and reduce the risk of false precision.

Nevertheless, evaluators may propose alternative response options. In the event that the respondent does not provide enough information to determine a counterfactual difference in energy savings, the recommended approach is to assign the midpoint value of 25. However, evaluators may also propose an alternative approach. The SWE requests that those evaluators provide their rationale for such alternatives.

C.6.3 Treatment of "Don't Know" Responses

As described above, in the case of "don't know" responses to one of the free ridership questions, the Common Method assigns the appropriate midpoint score. For example, if a respondent cannot provide any response to the main counterfactual question for the intention component, the method assigns the midpoint value of 25 for that component.

One objection raised was that assigning a midpoint value will inflate the free ridership estimate in cases where mean free ridership is less than 50%. For example, Controlling for Socially Acceptable Response Bias, showed a mean intention value of 16.8 for non-residential programs. If the midpoint value of 25, rather than the mean of 16.8, is substituted for a "don't know" response to the intention component, the resulting total free ridership value will be inflated.

A proposed alternative to imputing the mean of non-missing responses is to exclude cases with "don't know" responses and replace them with another. Both of those treatments may be problematic, as they assume that "don't know" responders are otherwise similar to the rest of the sample. However, the mere fact that they could not answer the intention counterfactually suggests they may differ from other respondents in some important respects that might affect their overall free ridership level. Generally, imputing the mean for missing responses is not considered best practice.[122]

In previous research, members of the SWE could not use the non-residential data described above to reliably investigate the question of whether "don't know" responders differ from others, as only three non-residential respondents (2% of the sample of 158) gave a "don't know" response to the intention question. However, in the residential dataset, 70 respondents (6% of the sample of 1,252) gave "don't know" responses.[123]

Previous members of the SWE therefore investigated whether respondents who had intention "don't know" responses differed from other respondents on the influence component of the free ridership battery. On average, respondents who gave an intention response (n = 1,164) indicated a maximum program influence of 4.4 on a 1-to-5 scale, while those who gave an intention "don't know" response (n = 70) indicated a maximum program influence of 4.1. This difference was marginally significant (F = 3.2, p = .07). While this finding does not conclusively show that "don't know" respondents differ from others, it argues against assuming no difference.

We recognize that imputing the midpoint may be considered arbitrary. Moreover, our experience is that "don't know" responses are infrequent, and so the way in which they are handled likely will not have a great impact on the resulting free ridership estimates. Evaluators may implement alternative approaches to handling "don't know" responses in addition to assigning the midpoint and report both

[122] Enders, C.K. *Applied Missing Data Analysis*, New York: The Guilford Press, 2010.
[123] The percentage of respondents who gave "don't know" responses to the *influence* component was even lower – 1% for both residential and non-residential samples. Similarly, in a dataset of 228 non-residential respondents from a different evaluation conducted in Ontario, 2% of respondents gave *intention* "don't know" responses and none gave *influence* "don't know" responses.

results. As an alternative approach, we recommend using linear regression to predict the intention score from each respondent's influence score.

C.6.4 Consistency Checks and Related Issue

Consistency checks are frequently used in social and epidemiological research, but there are reasons not to include consistency checks in a free ridership survey.

The assumption that the inconsistency can be resolved accurately may be unfounded. That assumption is based on the belief that the questioner can accurately and reliably determine which of two inconsistent responses is the correct one. A respondent confronted with inconsistent responses may seek to resolve the consistency, but that does not mean that the final response will be accurate. Instead, the response may be influenced by self-enhancement motivation.[124]

Other reasons not to confront respondents with inconsistent responses are that doing so may make respondents feel uncomfortable, and as a result, it could color later responses; it also lengthens the survey. Lengthening the survey, and perhaps even inducing some discomfort, may be acceptable if the result is better data. However, as argued above, there is reason to believe that it will not do so. Further, the need to assess which response is correct brings more evaluator subjectivity into the assessment. Therefore, we recommend against consistency checks.

C.6.5 Influence from Previous Program Years or Cycles

One evaluator asked whether influence to participate in a program that comes from participation in a previous year (or previous phase) is considered free ridership.

Our experience has been that most regulators limit consideration to the current year or phase. In practice, it may be difficult to determine whether program influence was from the current year or phase or from an earlier year or phase.

[124] Swann, William B., Jr. "Self-Verification Theory." In P. Van Lange, A.W. Kruglanski, and E.T. Higgins (eds.), *Handbook of Theories of Social Psychology*. Thousand Oaks, CA: Sage Publications, 2011.

Appendix D Common Approach for Measuring Spillover for Downstream Programs

D.1 Introduction

The PA PUC Implementation Order specifies that the NTG ratio for Phase V of Act 129 is to be treated in the same way as previous Phases. Specifically, for compliance purposes, the NTG ratios for Phase V programs continue to be set a 1.0 – basing compliance with energy and demand reduction targets on gross verified savings. However, the PUC order also states that the EDCs should continue to use net verified savings to inform program design and implementation.

The SWE recommends standardization – at a minimum within the EDCs' measurement activities and ideally across all EDCs – for provision of consistency in explaining program effects. The Framework also defines participant and non-participant spillover (spillover or SO) and recommends the consideration of trade ally surveys and reports for assessing the non-participant portion of a program's spillover impact. However, the SWE has determined that while estimation of non-participant spillover is desirable, it is not required. If assessed, non-participant spillover may be assessed through either a general population (non-participant) survey or through a survey of trade allies.

A description of a common approach for measuring free ridership for downstream programs is included in Appendix C. In it, we discuss the reasons for having a uniform NTG approach for the EDCs.

The following sections describe the draft common approach to the assessment of participant and non-participant spillover.

As is the case with the common approach to free ridership estimation, EDCs and their ECs may, if they wish, use alternative approaches in parallel with the common approach to assessing participant spillover through self-report surveys or add elements to the common approach, but they should be able to report results from the common approach as described below in addition to reporting results from alternative or modified approaches to assessing participant spillover. Moreover, EDCs and their ECs may propose alternative approaches for programs for which the common method may not be applicable, such as approaches focusing on midstream or upstream influences for non-participant spillover.

D.2 Sampling

The *Framework* does not specify confidence and precision levels for estimating spillover. The SWE recommends – but does not require – that the evaluation strive to achieve confidence and precision levels sufficient to provide meaningful feedback to EDCs.

As noted above, the SWE has determined that, while estimation of non-participant spillover is desirable, it is not required. If assessed, the sampling approach should produce a sample that is representative of the target population (non-participants or trade allies) or capable of producing results that can be made representative through appropriate weighting of data. In the case of trade ally surveys, the sampling plan should take trade ally size (e.g., total sales, total program savings) and type of equipment sold and installed (e.g., lighting or non-lighting) into consideration.

D.3 Participant Spillover

The following provides a description of the SWE's recommended approach for assessing participant spillover. It begins with an overview of the recommended approach. Following are detailed descriptions of the specific approaches for residential and non-residential participant spillover. The latter cover the SWE's recommended questions and response options to include in participant surveys as well as recommended computational rules for converting survey responses to inputs into the formulas for calculating spillover. The residential and non-residential participant surveys are slightly different.

D.3.1 Overview of Recommended Common Protocol

For both the residential and non-residential sectors, the participant spillover approach will assess the following for each participant:

- The number and description of non-incented energy efficiency measures taken since program participation.
    - This may include all energy efficiency measures, even if not eligible for program incentives. However, EDCs should distinguish between program-eligible and other types of measures (including measures that are in the TRM but not eligible for a specific program and energy efficient measures not in the TRM) in their analyses. See further discussion in Section D.3.2.
- An estimate of energy savings associated with those energy efficiency measures. (Details in Section D.3.2.)
- The program's influence on the participant's decision to take the identified measures, assessed with a rating scale and converted to a proportion, with possible values of 0, 0.5, and 1. (Details in Section D.3.2.)
- The likelihood the participant would have implemented the identified measures without program participation, assessed with a rating scale and converted to a proportion, with possible values of 0, 0.5, and 1. (Details in Section D.3.2.)

The specific methods for the residential and non-residential sector will differ somewhat in details of program influence assessment and estimation of the measure-specific energy savings.

As detailed below, evaluators will calculate spillover savings in four categories:

- For program-eligible measures.
- For measures in the TRM but not eligible for incentives for the program in question.
- For measures not in the TRM but for which the EDC's evaluator can provide reasonable documentation of savings.
- For all measures in any of the above categories.

For each of the above categories, the evaluators will complete the following tasks:

- Calculate total spillover savings for each participant as the sum of measure savings by number of units by influence score.
- Total the savings associated with each program participant, to give the overall participant SO savings.
- Multiply the mean participant SO savings for the participant sample by the total number of participants to yield an estimated total participant SO savings for the program.
- Divide that total savings by the total program savings to yield a participant spillover percentage.

D.3.2 Residential Participant Spillover – Detailed Methods

The residential participant spillover survey will include questions to assess, for each participant, the number and description of non-incented energy efficiency measures taken since program participation; and the program's influence on the participant's decision to take those measures.

Identification of Non-Rebated Residential Measures

The survey will assess the purchase and installation of any energy efficient measures, whether eligible for program rebates, in the TRM but not eligible, or not in the TRM. The survey will ask participants a series of questions similar to the following to determine whether they installed any additional energy efficient measures without receiving a rebate:

- You received a rebate for installing [list of rebated measures]. Since participating in the program, have you installed any additional [list of rebated measures] for which you did not receive a rebate?
    - [IF YES:] How many/how much have you installed?[125]
- Since participating in the program, have you installed any other energy efficient products or equipment, or made any energy efficiency improvements for which you did NOT receive a program rebate?
    - [IF YES:] What type of other energy efficient improvements, products, or equipment did you install? [Record description of each additional installed measure]
    - [FOR EACH MEASURE:] How many/how much did you install?

Assessment of Program Influence on Residential Measures

The survey will ask respondents about the level of influence the prior program participation had on their decision to install the additional measures. The survey may apply a single influence assessment to all measures, under the assumption that residential respondents are not likely to report different levels of

[125] Ask "how many" for unit items, such as lamps, appliances, and so forth. Ask "how much" for items installed by quantity, such as weather sealing or insulation.

program influence for different measures. At the evaluator's discretion, the survey may assess influence for each measure identified.

The SWE recommends that the influence question identify various ways in which the program participation might have influenced the decision to install additional measures. For example, evaluators may consider a question similar to the following:

- On a 1 to 5 scale, with 1 meaning "not at all influential" and 5 meaning "extremely influential," how influential were each of the following on your decision to [vary wording as appropriate:] install the additional equipment/product(s)/improvement(s)?126
  - Information about energy savings from utility marketing, program representatives, retailers, or contractors
  - Your satisfaction with the equipment for which you had received a rebate
  - Your installation of [rebated measure(s)] made you want to do more to save energy

Program influence is assessed as the maximum influence rating given to the four program elements.

- **Example**: A respondent gives influence ratings of 3, 5, and 3, respectively, energy savings information, satisfaction with equipment, and desire to do more. Therefore, the program influence rating is 5 because at least one program element was "extremely influential."

The maximum influence rating is assigned a value that determines what proportion of the relevant measures' savings is attributed to the program:

- A rating of 4 or 5 = 1.0 (full savings attributed to the program).
- A rating of 3 = 0.5 (half of the savings attributed to the program).
- A rating of 1 or 2 = 0 (no savings attributed to the program).

At the evaluator's discretion, to provide additional relevant feedback to the program, the survey may ask participants whether there was a reason that they did not receive an incentive for the additional energy efficient technologies.

Assessment of the Likelihood of Installing Residential Measures

The survey will ask respondents to assess the likelihood they would have installed the additional measures in the absence of program participation. The survey may apply a single likelihood assessment to all measures, based on the assumption that residential respondents are not likely to report different

[126] The survey should ask about all three of the above items, as they may have had differing levels of influence. Assessments of "overall program influence" may incorporate the lower ratings of some program elements. However, the final program influence rating will be the maximum influence of any single program element. Moreover, a single question about overall "program influence" may not incorporate influence from information that a program-influenced retailer or contractor provided and does not get at the possible cognitive processes that may have resulted from having undertaken program-induced energy savings.

levels of likelihood for different measures. At the evaluator's discretion, the survey may assess likelihood separately for each measure identified.

For example, evaluators may consider a question similar to the following: On a 1 to 5 scale, with 1 meaning "not at all likely" and 5 meaning "extremely likely," how likely is it that you would still have [vary wording as appropriate:] installed the additional equipment/product(s)/improvement(s) if you had not participated in the program?

The likelihood rating is assigned a value that determines what proportion of the relevant measures' savings is attributed to the program:

- A rating of 4 or 5 = 0 (no savings attributed to the program).
- A rating of 3 = 0.5 (half of the savings attributed to the program).
- A rating of 1 or 2 = 1 (full savings attributed to the program).

Assessment of Energy Savings for Residential Spillover

Where applicable, the savings for each additional measure installed will be calculated per the TRM for a rebated measure installed through the program. For partially deemed measures, the SWE and EDCs/EDC evaluators will develop conservative working assumptions for any required inputs (e.g., square footage of home, R-value improvement, replaced wattage) or may identify average verified savings for such measures.

For measures not in the TRM, the evaluator should identify the source and methodology used to assess per-item savings.

Calculation of Total Residential Spillover and Savings Rate

Evaluators will calculate summed spillover savings in four categories:

- For program-eligible measures.
- For measures in the TRM but not eligible for incentives for the program in question.
- For measures not in the TRM but for which the EDC's evaluator can provide reasonable documentation of savings.
- For all measures in any of the above categories.

Evaluators will first calculate spillover savings for each spillover measure reported as the product of the measure savings, number of units, and the average of the influence and likelihood scores:

$$Measure\ SO = Measure\ Savings * Number\ of\ Units * \left[\frac{Program\ Influence + Likelihood}{2}\right]$$

For each of the above categories, the evaluators then will complete the following tasks:

- Total the savings associated with each program participant, to give the overall participant SO savings.

$$Participant\ SO = \Sigma Measure\ SO$$

- Multiply the mean participant SO savings for the participant sample by the total number of participants to yield an estimated total participant SO savings for the program.

$$\Sigma Participant\ SO\ (population) = \frac{\sum Participant\ SO\ (sample)}{Sample\ n} \times \text{Population}\ N$$

- Divide that total savings by the total program savings to yield a participant spillover percentage:

$$\%\ Participant\ SO\ = \frac{\sum Participant\ SO\ (population)}{Program\ Savings} \times 100$$

D.3.3 Non-Residential Participant Spillover – Detailed Methods

The participant spillover survey includes questions to assess, for each participant, the number and description of non-incented energy efficiency measures taken since program participation; and the program's influence on the participant's decision to take those measures. The approach for non-residential participant spillover is similar to that for residential but differs in some details.

Identification of Non-Rebated Non-Residential Measures

The survey will assess the purchase and installation of any energy efficient measures, using questions similar to the following:

- Since your participation in the program, did you install any ADDITIONAL energy efficiency products or equipment, or made any energy efficiency improvements that did NOT receive incentives through any utility program?
    - [IF YES:] Please describe the energy efficiency equipment installed or energy efficiency improvement? [Probe for measure type, size, and quantity]

The questioner should attempt to document all additional, non-rebated equipment installed since program participation, whether eligible for program rebates, in the TRM but not eligible, or not in the TRM.

Assessment of Program Influence on Non-Residential Measures

The survey will ask respondents about the level of influence the prior program participation had on their decision to install the additional measures. For example, evaluators may consider a question similar to the following:

- On a 1 to 5 scale, with 1 meaning "not at all influential" and 5 meaning "extremely influential," how influential was your participation in the [NAME OF PROGRAM] on your decision to [vary wording as appropriate:] install the additional equipment/complete the energy efficiency improvement(s)?

At the evaluators' discretion, the survey may ask the above influence question only once to cover all additional energy efficient installations or improvements or separately for different energy efficient installations or improvements. In the event that a respondent reports many (e.g., more than three) additional non-rebated measures, evaluators have the option of assessing influence for some of them (e.g., the three that deliver the greatest energy savings) and assigning the mean influence score from those measures to the remaining ones.

For each additional energy efficient installation or improvement, the influence rating is assigned a value that determines what proportion of the measure's savings are attributed to the program:

- A rating of 4 or 5 = 1.0 (full savings attributed to the program).
- A rating of 3 = 0.5 (half of the savings attributed to the program).
- A rating of 1 or 2 = 0 (no savings attributed to the program).

At the evaluator's discretion, to provide additional relevant feedback to the program, the survey may ask participants whether there was a reason that they did not receive an incentive for the additional energy efficient technologies.

Assessment of the Likelihood of Installing Non-Residential Measures

The survey will ask respondents to assess the likelihood they would have installed the additional measures in the absence of program participation. For example, evaluators may consider a question similar to the following:

- On a 1 to 5 scale, with 1 meaning "not at all likely" and 5 meaning "extremely likely," how likely is it that you would still have [vary wording as appropriate:] installed the additional equipment/product(s)/improvement(s) if you had not participated in the program?

At the evaluators' discretion, the survey may ask the above likelihood question only once to cover all additional energy efficient installations or improvements or separately for different energy efficient installations or improvements. In the event that a respondent reports many (e.g., more than three) additional non-rebated measures, evaluators have the option of assessing likelihood of installation for some of them (e.g., the three that deliver the greatest energy savings) and assigning the mean influence score from those measures to the remaining ones.

For each additional energy efficient installation or improvement, the influence rating is assigned a value that determines what proportion of the measure's savings are attributed to the program:

- A rating of 4 or 5 = 0 (no savings attributed to the program).
- A rating of 3 = 0.5 (half of the savings attributed to the program).
- A rating of 1 or 2 = 1 (full savings attributed to the program).

Assessment of Energy Savings

Where applicable, the savings for each additional measure installed will be calculated per the TRM for a rebated measure installed through the program. For partially deemed measures, the SWE and EDCs/EDC evaluators will develop conservative working assumptions for any required inputs (e.g., square footage of home, R-value improvement, replaced wattage) or may identify average verified savings for such measures.

For measures not in the TRM, the evaluator may conduct a brief engineering analysis to assess savings or, if applicable, identify an alternative source and methodology for assessing savings.

Calculation of Total Non-Residential Spillover and Savings Rate

The calculation of non-residential spillover and savings rate is essentially the same as for residential.

Evaluators will calculate summed spillover savings in four categories:

- For program-eligible measures.
- For measures in the TRM but not eligible for incentives for the program in question.
- For measures not in the TRM but for which the EDC's evaluator can provide reasonable documentation of savings.
- For all measures in any of the above categories.

Evaluators will first calculate spillover savings for each spillover measure reported as the product of the measure savings, number of units, and the average of the influence and likelihood scores:

$$Measure\ SO = Measure\ Savings * Number\ of\ Units * \left[\frac{Program\ Influence + Likelihood}{2}\right]$$

For each of the above categories, the evaluators then will complete the following tasks:

- Total the savings associated with each program participant, to give the overall participant SO savings.

$$Participant\ SO = \Sigma Measure\ SO$$

- Multiply the mean participant SO savings for the participant sample by the total number of participants to yield an estimated total participant SO savings for the program.

$$\Sigma Participant\ SO\ (population) = \frac{\sum Participant\ SO\ (sample)}{Sample\ n}$$

- Divide that total savings by the total program savings to yield a participant spillover percentage:

$$\%\ Participant\ SO = \frac{\sum Participant\ SO\ (population)}{Program\ Savings}$$

D.4 Non-Participant and Total Spillover

The SWE has determined that while estimation of non-participant spillover is desirable, it is not required. Non-participant spillover may be assessed either through a general population (non-participant) survey or through a survey of trade allies.

D.4.1 Non-Participant Survey

If a general population survey is selected, it should assess the following for each survey respondent:

- The number and description of non-incented energy efficiency measures taken in the program period.
- An estimate of energy savings associated with those energy efficiency measures.
- The program's influence on the non-participant's decision to take the identified measures, assessed with a rating scale and converted to a proportion, with possible values of 0, 0.5, and 1.
- The likelihood the non-participant would have implemented the identified measures in the absence of the program, assessed with a rating scale and converted to a proportion, with possible values of 0, 0.5, and 1.

Evaluators should submit draft survey questions to the SWE.

D.4.2 Trade Ally Survey

The following provides an overview of the SWE's recommended approach to assessing spillover through a trade ally survey, followed by the SWE's recommended questions and response options to include in participant and trade ally surveys to assess residential and non-residential SO as well as recommended computational rules for converting survey responses to inputs to the formulas for calculating SO, described above. The residential and non-residential participant surveys are slightly different and are described in separate subsections. The residential and non-residential trade ally surveys are essentially identical and are described in a single subsection.

Overview of Recommended Trade Ally Approach

If an evaluator chooses to assess non-participant spillover through trade ally surveys, separate surveys should be conducted for the residential and non-residential sectors. Each survey should assess the following for each sampled respondent:

- The number of program-qualified measures sold or installed within the specified sector, in the specified utility's service territory, in the specified program year.
- The percentage of such installations that received rebates from the specified program.
- The trade ally's estimate of the proportion of their sales or installations of non-rebated measures that went to prior program participants.
- The trade ally's judgment of the specified program's influence on sales of the common program-qualified but not rebated measures, assessed with a rating scale and converted to a proportion, with a minimum value of 0 and a maximum value of 1.

The survey should estimate total sales of all program-qualified measures by asking trade allies to report sales of their most commonly sold program-qualifying measures and determining what proportion of their total sales of high-efficiency products those measures made up (detailed below). Trade ally survey questions should ask about sales within a specific sector (residential or non-residential). If an evaluation plan calls for a single trade ally survey in a given sector to provide SO figures across multiple programs within that sector, that survey should be worded to ensure that the trade ally understands that responses should refer to the multiple programs.

Identification of Non-Rebated Measures

The trade ally surveys will ask about sales or installations of the program's most common qualified measures. Theoretically, the survey should assess sales or installations of all program-qualified measures. Otherwise, it will undercount SO. However, doing so would create unreasonable burden on the respondents and would not likely produce reliable results. Therefore, the recommended common method takes the following approach.

First, evaluators should identify each sampled trade ally's most rebated measures as well as other commonly rebated program measures of the type pertinent to the trade ally.

The survey should assess the number of non-rebated units sold of each of the respondent's most rebated measures within the territory of the EDC in question. The introduction to the survey should make it clear to respondents that questions about sales of measures pertain to measures sold within that EDC's territory and that responses should refer to a given sector (residential or non-residential) and to all that EDC's applicable programs within that sector.

To prevent undue burden, the survey should restrict the number of measures investigated to no more than four. For each of those measures, the survey should ask respondents questions like the following:

- During the program year, how many [measure] did you sell/install within the service territory of [EDC]?
- Approximately what percentage of your [measure] installations in [EDC] service territory received rebates through the program?

By subtraction, the response to Question 2 provides the percentage of non-rebated units, of a specific type, sold/installed.

For each of the respondent's most sold program-rebated measures, the number of non-rebated units will be estimated as the total number of units sold/installed multiplied by the non-rebated percentage.

As indicated above, it is impractical for the survey to attempt to estimate the number of units of all program-qualified measures that a respondent sold. This means that the above procedure will underestimate spillover. As a way of providing some information on the possible degree to which spillover is underestimated, the survey should ask respondents to estimate the percentage that their most commonly rebated products, combined, comprise of their total sales/installations of high-efficiency products, using a question like the following:

- Thinking about those types of products together, what percentage do they make up of your total dollar sales of high-efficiency products?

The purpose of this question is not to inform a precise and reliable estimate of additional spillover, but rather to provide information on the possible degree to which spillover is underestimated.

Assessment of Program Influence

For each of the identified measures, the survey will ask respondents about the level of influence the program had on their sales/installations of non-rebated program-qualified measures, using a question like the following:

- Using a 1 to 5 likelihood scale, where 1 is "not at all influential" and 5 is "extremely influential," how influential was the program on your sales of non-rebated high-efficiency products of that type to your customers?

For each measure identified, the maximum influence rating is assigned a value that determines what proportion of the measure's savings is attributed to the program:

- A rating of 4 or 5 = 1.0 (full savings attributed to the program).
- A rating of 3 = 0.5 (half of the savings attributed to the program).
- A rating of 1 or 2 = 0 (no savings attributed to the program).

Assessment of Energy Savings

The savings for each additional measure installed will be calculated per the TRM for a rebated measure installed through the program. For partially deemed measures, the SWE and EDCs/EDC evaluators will develop conservative working assumptions for any required inputs (e.g., square footage of home, R-value improvement, replaced wattage) or may identify average verified savings for such measures.

Calculation of Trade-Ally-Reported Spillover (SO)

For each surveyed trade ally, the total SO of each reported measure (i.e., the commonly rebated measures) will be calculated as follows:

$$Reported\ Measure\ SO = Measure\ Savings * Number\ of\ Units * Program\ Influence$$

The SO from each measure will be summed for each surveyed trade ally to calculate the total SO for that trade ally. Total trade-ally-reported SO for a program can be estimated one of two ways:

- Calculate the mean total SO per trade ally and multiply it by the total number of trade allies, if known, to estimate total SO for the program.
- Calculate the mean SO percentage for each sampled trade ally as the trade ally's total SO divided by the trade ally's total program savings; calculate the mean SO percentage across sampled trade allies (weighted by trade ally size; see below) and multiply that mean SO percentage by the total program savings (from the program database) to estimate total SO for the program.

In either case, the mean total SO or mean SO percentage for trade ally-reported measures should be weighted by trade ally size using total program sales of non-rebated high-efficiency equipment (if available) or by a reasonable proxy, such as total program incentives. The means also should be weighted by trade ally type (e.g., lighting or non-lighting).

Total trade-ally-reported SO can be divided by the total program savings to yield a total SO percentage, as:

$$\%\ Total\ Trade\ Ally\ (TA) Reported\ SO\ = \frac{\sum Total\ TA\ Reported\ SO\ Across\ all\ Program\ TAs}{Program\ Savings}$$

The evaluators should calculate and report the weighted mean percentage of total sales of high-efficiency equipment that the reported SO measures constitute. The percentage should be weighted by total sales of high-efficiency equipment (if available) or by a reasonable proxy, such as total program incentives. (Again, the purpose is not to yield a precise and reliable estimate of additional spillover, but to provide a best available indication of the degree to which spillover may be undercounted.)

Total and Non-Participant Spillover

The above approach theoretically yields (but underestimates) total SO because it does not differentiate between sales of non-rebated measures to program participants and non-participants.

If responses to the trade ally survey indicate that the trade-ally-identified commonly sold program-rebated measures comprise a large percentage (e.g., 90% or more) of all high-efficiency equipment sold, then evaluators should attempt to determine what percentage of the total trade-ally-identified SO is from non-participants by subtracting the total participant SO for that sector from the total trade-ally-reported SO, as follows:

$$\sum Nonparticipant\ SO = \sum Total\ TA\ Reported\ SO - \sum Participant\ SO$$

That total, divided by the total program savings, yields a non-participant SO percentage, as:

$$\%\ Nonparticipant\ SO\ = \frac{\sum Nonparticipant\ SO}{Program\ Savings}$$

If the trade-ally-identified commonly sold program-rebated measures do not comprise a large percentage (e.g., 90% or more) of all high-efficiency equipment sold, then subtracting participant SO likely will not yield an accurate estimate of non-participant SO. In that case, evaluators should report the total trade-ally-reported SO and participant SO.