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

FINAL VERSION

May 8, 2018

CONTRACTED UNDER THE PENNSYLVANIA PUBLIC UTILITY COMMISSION'S RFP 2015-3 FOR THE STATEWIDE EVALUATOR

PREPARED BY the Statewide Evaluation Team:
NMR Group, Inc.
EcoMetric Consulting, LLC
Demand Side Analytics, LLC

Table of Contents

Figures

Tables

Equations

Prepared By

Salil Gogte – EcoMetric Consulting, LLC

Greg Clendenning, Rohit Vaidya, Lynn Hoefgen – NMR Group, Inc.

Jesse Smith – Demand Side Analytics, LLC

Acknowledgments

This Phase III Evaluation Framework builds off the Phase II Evaluation Framework and updates it for Phase III. The SWE would like to acknowledge the hard work of GDS Associates, Research Into Action, and Apex Analytics in preparing the previous versions of the Framework.

List of Acronyms

B/C Ratio: Benefit/Cost Ratio

BTUh: BTU-hours

CBL: Customer Baseline

CDO: Commercial Date of Operation

CEEP: Conservation, Economics, and Energy Planning [now called the Bureau of Technical Utility Services (TUS)]

CFL: Compact Fluorescent Light

CPITD: Cumulative Program Inception to Date

Cv: Coefficient of Variation

DLC: Direct Load Control

DR: Demand Response

DSM: Demand Side Management

EC: Evaluation Contractor

ECM: Energy Conservation Measure

EDC: Electric Distribution Company

EE: Energy Efficiency

EE&C Plan: Energy Efficiency and Conservation Plan

EER: Energy-Efficiency Ratio

EISA: Energy Independence and Security Act of 2007

ELRP: Emergency Load Reduction Program

EM&V: Evaluation, Measurement, and Verification

FPC: Finite Population Correction Factor

HIM: High-Impact Measure

HVAC: Heating, Ventilating, and Air Conditioning

ICSP: Implementation Conservation Service Provider

IMP: Interim Measure Protocol

IPMVP: International Performance Measurement and Verification Protocol

ISD: In-Service Date

I-SIR: Independent Site Inspection Reports

kW: Kilowatt

kWh: Kilowatt-Hour

LED: Light-Emitting Diode

MEP: Measure-specific Evaluation Protocol

MPI: Market Progress Indicator

M&V: Measurement and Verification

NPV: Net Present Value

NTG: Net-to-Gross Savings

NTGR: Net-to-Gross Ratio

PEG: Program Evaluation Group

PJM: PJM Interconnection, LLC

PUC: Pennsylvania Public Utility Commission

PY: Program Year

RA-SIR: Ride Along Site Inspection Report

SEM: Simple Engineering Model

SSMVP: Site-Specific M&V Plan

SWE: Statewide Evaluator

SWE Team: Statewide Evaluation Team

TOU: Time-of-Use

TRC: Total Resource Cost Test

TRM: Technical Reference Manual

TUS: Bureau of Technical Utility Services [formerly the Conservation, Economics, and Energy Planning (CEEP)]

TWG: Technical Working Group

UMP: Uniform Methods Project

VOI: Value of Information

VFD: Variable Frequency Drive

Appendix A contains a glossary of terms.

Section 1 Introduction and Purpose of the Evaluation Framework

This Evaluation Framework includes guidelines and expectations for the seven 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 and objectives of Act 129. The EDCs are Duquesne Light Company, Metropolitan Edison Company, PECO Energy Company, Pennsylvania Electric Company, Pennsylvania Power Company, PPL Electric Utilities Corporation, and West Penn Power Company.

Through a Request for Proposal (RFP) process, the PUC contracted with a Statewide Evaluation (SWE) Team for all Phases of Act 129. The SWE Team's objective is to complete a comprehensive evaluation of the Act 129 EE&C programs implemented by the seven EDCs in Pennsylvania.

The SWE Team proposed a scope of work that met all of the requirements for tasks and deliverables in the PUC's RFP, including the level of verification described in the RFP and at the pre-bid meeting. The approach involves auditing verifications completed by EDC evaluators.

To conduct these activities, the SWE Team will collaborate with the seven EDCs, their evaluation teams, and the PUC staff in order to develop appropriate, effective, and uniform procedures to ensure that the performance of each EDC's EE&C programs is verifiable and reliable and meets the objectives of the Act 129 under which the programs were developed.

In accordance with the RFP and the scope of work for the Statewide Evaluator, the SWE Team's tasks are as follows:

- 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 to confirm the accuracy of EDC-reported and 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)
  - Statewide Baseline Study to characterize the market and assess equipment saturation and energy efficiency levels
  - Statewide Market Potential Study to provide estimates to inform PUC decisions regarding additional electric energy and load reductions for Phase IV 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 to the EDCs and their evaluation

contractors. While the document is not a Commission Order, and therefore not mandatory, EDCs that align their EM&V processes with the Evaluation Framework should expect less scrutiny from the SWE 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 among a Program Evaluation Group (PEG) of the EDCs, their evaluators, the SWE Team, and TUS. During these discussions, the PEG will clarify and interpret the TRM, recommend additional measures to be included in the TRM, and define guidelines for acceptable measurement protocols for custom measures in order to mitigate evaluation risks to the EDCs. This will require clear and auditable definitions of kWh/yr 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 on claiming savings
- Guidance and requirements on gross impact evaluation procedures
- Guidance and requirements on process evaluation procedures
- Guidance and requirements on net-to-gross (NTG) analysis
- Guidance and requirements on cost-effectiveness analysis
- Guidance and requirements on statistics and confidence/precision
- Required reporting formats
- Data management and quality control guidelines and requirements
- Guidance and requirements on 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 evaluation contractor and assess individual and collective EDC progress toward attainment of Act 129 energy savings targets
- Criteria the SWE Team will use to review and assess EDC evaluations

Per the PUC, the EDCs must adopt and implement the approved Evaluation Framework upon its release. 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 will provide PUC-approved updates as addenda to the Evaluation Framework.

1.1 Act 129 Requirements for the Statewide Evaluation

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

1. Developing an Evaluation Framework
2. Monitoring and verifying EDC data collection
3. Developing and implementing quality assurance processes

4. Defining performance measures by customer rate class (e.g., sector)

The SWE 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 IV of Act 129. The audits will include an analysis of each EDC plan from process, impact, and cost-effectiveness standpoints. The annual audits will include an analysis of plan and program impacts (energy and demand savings) and cost-effectiveness. The SWE is to report results and provide recommendations for plan and program improvements. The RFP states that the SWE will produce an accurate assessment of the potential for energy efficiency and demand response 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 III Implementation Order and Act 129.[1]

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

- **Impact evaluations** seek to *quantify t*he energy, 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 and to characterize their efficiency and effectiveness.
- **Cost-effectiveness tests** seek to *assess* whether the avoided monetary cost of supplying electricity is greater than the monetary cost of energy efficiency 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.

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 deliverable, per these categories:

- Statewide Studies
- Audit and Assess EDC Phase III Programs and Results

[1] The PUC has been charged by the Pennsylvania General Assembly pursuant to Act 129 of 2008 ("Act 129") with establishing an Energy Efficiency and Conservation (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 11, 2015, the PUC entered an Implementation Order at Docket No. M-2014-2424864. As part of the Implementation Order and Act 129, the PUC issued an RFP for a Statewide Evaluator (on November 23, 2015) to evaluate the EDCs' Phase III EE&C programs.

- Databases
- Primary Data Collection and Impact Analyses
- EDC Plan Review
- Reporting (Semi-Annual and Annual)
- Best Practices
- Other

When appropriate, the SWE has classified tasks within the EDCs' primary responsibilities as a role of the implementation conservation service provider(s) (ICSP) or evaluation contractor (EC).

Table 1: Roles and Responsibilities - Statewide Studies

Task and/or Deliverable	EDC	SWE	PUC
Conduct energy efficiency baseline studies to support Market Potential Study		XX	
Conduct electric energy efficiency Market Potential Study for targets to be achieved in a potential Phase IV EE&C Program		XX	
Conduct a Demand Response Potential Study for targets to be achieved in a potential Phase IV Demand Response Program		XX	
Review and get approval of Statewide Baseline and Market Potential Studies (the SWE would get approval of these studies from the Commission)			XX
Initiate and coordinate updates to TRM and interim updates (new protocols)		XX	
Approve TRM updates			XX
Initiate, scope, and conduct/coordinate statewide site inspections, statewide evaluation studies, review of data/studies from PA and other states to determine if the PA TRM appropriately estimates savings and/or to revise PA TRM protocols		XX	
Develop and conduct EDC-specific or broader studies and research such as NTG, program design best practices, and market effects studies	XX		
Coordinate the development of and approve the methodologies for EDC NTG, process evaluation, and market effects studies consistent with this evaluation framework		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 evaluation contractors		XX	XX
Review and update the Evaluation Framework		XX	
Approve the statewide Evaluation Framework and revisions			XX
Conduct impact evaluation, process evaluation, NTG analysis, and cost-effectiveness evaluation	EC		
Review/audit all EDC evaluation results, impact evaluation, process evaluation, NTG analysis, and cost-effectiveness evaluation		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 secure SharePoint site for maintenance and exchange of confidential data and information with EDCs		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 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		
Determine ex post verification of installation, measure operability, and energy savings	EC		
Analyze and document project, program, and portfolio gross and net energy and demand savings	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 EDC evaluator contractor performance of 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 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	

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 demand response program and portfolio net and gross impacts, as applicable, as well as cost-effectiveness and EDC progress in reaching targets; conduct process evaluation	EC		
Develop the statewide semi-annual and annual report templates; review EDC reports and advise the PUC of program and portfolio results: net and gross 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 annual reports			XX
Review EDC semi-annual and annual reports and SWE's semi-annual and annual reports on Act 129 programs: net and gross 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
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	
Organize and conduct periodic stakeholder meetings on evaluation results of EE and DR programs and proposed changes to the TRM		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 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 the most successful and why? • Which programs were the 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 in order to meet statutory targets?
Pennsylvania PUC	• What level of program energy savings was verified for each EDC and how does this compare to planning estimates and savings reported in EDC semi-annual and annual reports? • What assumptions related to energy and demand savings need to be updated in the future TRM versions? • What were the largest sources of uncertainty identified by EDC evaluators related to energy and demand savings and cost-effectiveness?	• Why did planning estimates and reported gross savings differ from verified gross savings? • Considering differences in planning estimates, 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' evaluation contractors?

Target Audience	Impact Questions	Process Questions
		• What were the process-related findings of all of the site inspections conducted by EDCs to verify equipment installation?
Pennsylvania EDCs	• What factors contributed to differences between planning estimates 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 estimates, reported gross savings, and verified gross savings? • What changes can the EDCs adopt to influence customer awareness, satisfaction, and adoption of EE&C programs?

Section 2 Policy Requirements

2.1 Requirements From the Phase III Implementation Order

Act 129 requires the PUC to establish an energy efficiency and conservation program that includes the following characteristics:

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

Further, the Phase I implementation order detailed that the PUC is responsible for "establishing the standards each plan must meet and providing guidance on the procedures to be followed for submittal, review and approval of all aspects of EDC energy efficiency and conservation (EE&C) plans."[7] Based on findings from the Phase II Market Potential Study dated February 2015, the PUC determined that the benefits of a Phase III Act 129 program would exceed its costs, and therefore adopted additional required incremental reductions in consumption and peak demand for another EE&C Program term of June 1, 2016, through May 31, 2021 (program years eight, nine, ten, eleven, and twelve). In its Phase III Implementation Order, the PUC established targets for those incremental reductions in electricity consumption for each of the seven EDCs in Pennsylvania; established demand response targets for six of the seven EDCs; 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 EDC EE&C plans for Phase III.[8]

2.1.1 Phase III Energy Reduction Targets for Each EDC

The PUC's June 2015 Implementation Order explained that it was required to establish electric energy consumption reduction compliance targets for Phase III of Act 129. Table 10

[4] This Act 129 requirement does not apply to an electric distribution company 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 page 4, at Docket No. M-2014-2424864, (*Phase III Implementation Order*), entered June 11, 2015,
[8] Pennsylvania Public Utility Commission, *Energy Efficiency and Conservation Program* Implementation Order, at Docket No. M-2014-2424864, (*Phase III Implementation Order*), entered June 11, 2015.

contains these targets as percentages and five-year cumulative totals in MWh/year for each of the seven EDCs.

Table 10: Act 129 Phase III Five-Year Energy Efficiency Reduction Compliance Targets

EDC	Portfolio EE Budget Allocation (Million $)	Program Acquisition Costs ($/1st-YR MWh Saved)	Five-Year Value of Reductions (MWh)	% of 2010 Forecast
Duquesne	$88.0	$199.5	440,916	3.1%
FE: Met-Ed	$114.4	$190.9	599,352	4.0%
FE: Penelec	$114.9	$202.9	566,168	3.9%
FE: Penn Power	$30.0	$190.4	157,371	3.3%
FE: West Penn	$106.0	$196.0	540,986	2.6%
PECO	$384.3	$195.8	1,962,659	5.0%
PPL	$292.1	$202.4	1,443,035	3.8%
Statewide	**$1,129.6**	**$197.8**	**5,710,488**	**3.9%**

The final Phase III Implementation Order also established demand response targets for each EDC covered by Act 129 (with no DR target for Penelec). The percentage reduction targets, as well as the value of reductions in MW, are reported in Table 11. It is important to note that the EDCs are not required to obtain peak demand reductions in the first program year of Phase III (PY8). The targets reported in Table 11 are for the other four program years in Phase III.

Table 11: Act 129 Phase III Five-Year Energy Demand Response Reduction Compliance Targets

EDC	5-Year DR Spending Allocation (Million $)	Program Acquisition Costs ($/MW/year)	Average Annual Potential Savings (MW)	% Reduction (Relative to 2007-2008 Peak Demand)
Duquesne	$9.77	$57,976	42	1.7%
FE: Met-Ed	$9.95	$51,210	49	1.8%
FE: Penelec	$0.00	$50,782	0	0.0%
FE: Penn Power	$3.33	$49,349	17	1.7%
FE: West Penn	$11.78	$46,203	64	1.8%
PECO	$42.70	$66,370	161	2.0%
PPL	$15.38	$41,622	92	1.4%
Statewide	**$92.90**	**$54,714**	**424**	**1.6%**

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

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

1. EDCs must include in their filing an EE&C Plan that obtains at least 3.5% of all consumption reduction requirements from the federal, state, and local governments, including municipalities, school districts, institutions of higher education, and nonprofit entities.
2. Each EDC Phase III EE&C Plan must obtain at least 5.5% of its consumption reduction requirements 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. More details about the low-income targets and requirements are provided below in Section 2.1.7. Act 129 also includes legislative requirements to include a number of energy efficiency measures for households at or below 150% of the federal poverty income guidelines that is proportionate to each EDC's total low-income consumption relative to the total energy usage in the service territory. The SWE 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.
3. EDCs will be awarded credit for all new, first-year, incremental savings delivered in each year of the Phase (rather than focusing on a cumulative approach, as was done in Phase II).
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 energy efficiency programs, subject to PUC approval. The PUC expects the EDCs to provide a reasonable mix of energy efficiency programs for all customers. However, each EDC's Phase III EE&C Plan must ensure that the utility offers each customer class at least one energy efficiency program.
7. Demand response programs will meet the following criteria:
    a. The EDCs will obtain no less than 85% of the target in any one event.
    b. Curtailment events shall be limited to the months of June through September.
    c. Curtailment events shall be called for the first six days that a peak hour of PJM's day-ahead forecast for the PJM RTO is greater than 96% of the PJM RTO summer peak demand forecast for the months of June through September for each year of the program.
    d. Each curtailment event shall last four consecutive hours.
    e. Each curtailment event shall be called such that it will occur during the day's forecasted highest peak hour above 96% of PJM's RTO summer peak demand forecast.

f. Once six curtailment events have been called in a program year, the peak demand reduction program shall be suspended for that program year.
g. The reductions attributable to a four-consecutive-hour curtailment event will be based on the average MW reduction achieved during each hour of an event.
h. Compliance will be determined based on the average MW reductions achieved from events called in the last four years of the program.
i. The EDCs, in their plans, must demonstrate that the cost to acquire MWs from customers who participate in PJM's ELRP is no more than half the cost to acquire MWs from customers in the same rate class that are not participating in PJM's ELRP. In addition, EDCs' DR programs are to allow for dual participation in Act 129 and PJM's ELRP; dual enrolled participants will have a 50% discount on Act 129 DR incentives imposed.

2.1.3 Carryover Savings from Phase II

The PUC's June 2015 Implementation Order for Phase III specifies that the EDCs are allowed to use savings attained in Phase II in excess of their targets for application toward Phase III targets. These carryover savings may only be savings actually attained in Phase II. In addition, the EDCs will only be allowed to carry over excess low-income savings into Phase III based on an allocation factor determined by the ratio of savings from low-income-specific programs. The allocation factor will be based upon the percent of verified low-income savings attributable to low-income-specific programs at the end of Phase II. For example, if an EDC has low-income savings in excess of their target in Phase II, and 40% of that EDC's verified low-income savings are attributable to low-income-specific programs, then the EDC may apply 40% of any excess low-income savings toward the Phase III 5.5% carve-out (Equation 1).[9]

Equation 1: Phase II Carryover Savings, Verified Low-Income (LI) Savings

$$Carryover_{PhIIVLIS} = \left(\frac{Verified\ Savings_{PhIILISP}}{Total\ Verified\ LI\ Savings, Ph\ II}\right) * Excess\ Verified\ LI\ Savings, Ph\ II$$

Where:

$Carryover_{PhIIVLIS}$ = Carryover, Phase II Verified Low-Income Savings
$Verified\ Savings_{PhIILISP}$ = Verified Savings, Phase II Low-Income-Specific Program

[9] Qualifying low-income savings from multifamily housing may be counted toward the low-income-specific savings, as well as savings from any program that was directly targeted to low-income customers. This includes all weatherization programs, energy efficiency kits and home energy report programs, and specifically targeted compact fluorescent lighting (CFL) and light-emitting diode (LED) lighting giveaway programs.

2.1.4 Incremental Annual Accounting

EDCs will be awarded credit for all new, first-year, incremental savings delivered in each year of the Phase. 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 or not a measure reaches the end of its EUL before the end of the phase does not impact compliance savings.

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

The PUC's Phase III 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 evaluation contractors should conduct NTG research in Phase III and how the results of this research can be incorporated into program planning.

2.1.6 Semi-Annual Reporting for Phase III of Act 129

For Phase III of Act 129, the EDC reporting requirements have been changed from quarterly to semi-annual. The EDCs are to submit, by January 15 of each year, a semi-annual report regarding the first six months of the program year. By July 15, the EDCs would submit a preliminary annual report for the program year that outlines the reported savings for that program year. Lastly, the EDCs would submit final annual reports by November 15 with gross verified savings for the program year, a cost-effectiveness evaluation (TRC Test), process evaluations, as well as items required by Act 129 and Commission orders. Section 4.1 provides more details.

2.1.7 Low-income Customer Savings

As noted earlier in Section 2.1.2, each EDC Phase III EE&C Plan must obtain at least 5.5% of its consumption reduction requirements 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. Low-income customers are defined as households whose incomes are at or below 150% of the Federal Poverty Income Guideline. As noted earlier in Section 2.1, low-income carryover for Phase III will be based on an allocation factor determined by the ratio of savings from low-income-specific programs to savings from non-low-income programs at the end of Phase II.

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 is proportionate to each EDC's total low-income consumption relative to the total energy usage in the service territory. The SWE 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 Energy Efficiency and Conservation (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 electric distribution company.
> (2) Energy efficiency and conservation measures shall include solar or solar photovoltaic 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 staff proposes 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 compact fluorescent light bulbs into multiple measures based on wattage. 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.

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

> (G) 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 12.

Table 12: Act 129 Phase III Low-income Carve-out Information

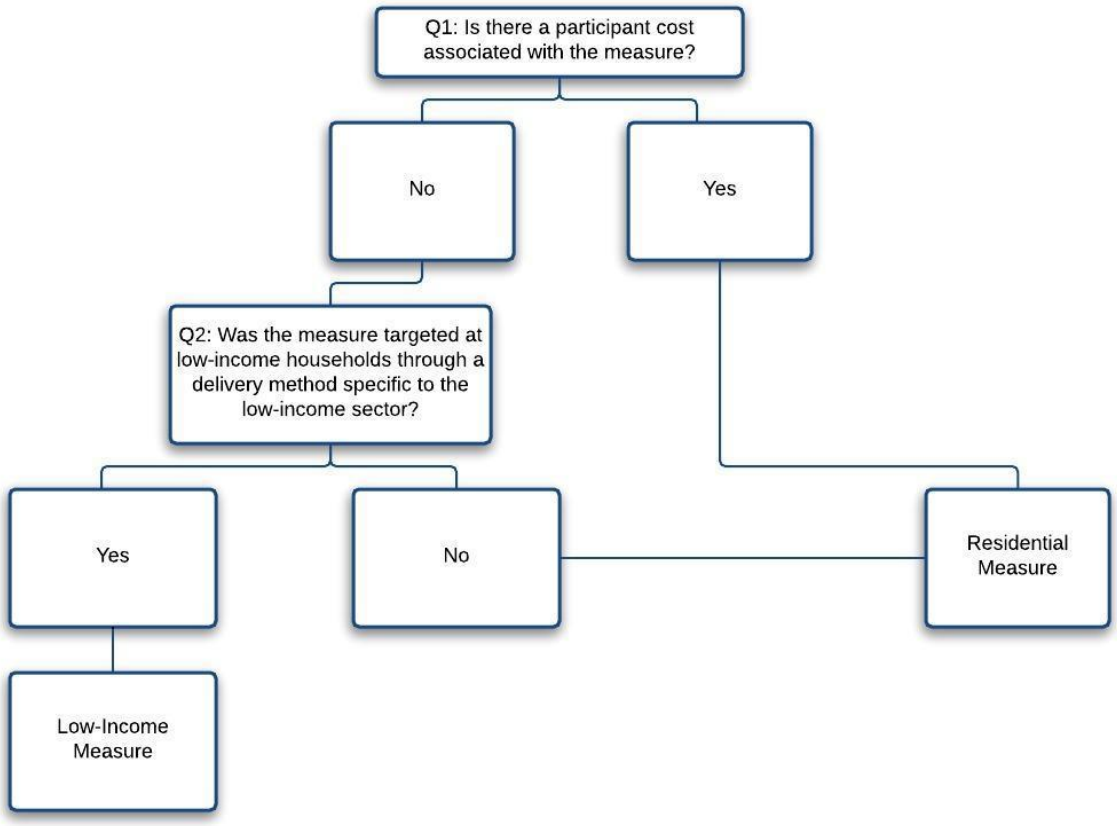
EDC	Proportionate Number of Measures	2016-2021 Potential Savings (MWh)	5.5% Low-Income Savings Target (MWh)
Duquesne	8.40%	470,609	24,250
FE: Met-Ed	8.79%	627,814	32,964
FE: Penelec	10.23%	598,612	31,139
FE: Penn Power	10.64%	170,182	8,655
FE: West Penn	8.79%	585,807	29,754
PECO	8.80%	2,080,553	107,946
PPL	9.95%	1,590,264	79,367

Please note that our proposed definition does *not* require that the measure/measure type be installed in order 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, mattress) but no customers end up having the measure (mattress) installed, it would still count toward satisfying the "proportionate measures" requirement.

The staff 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 an 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 2 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 Phase I and Phase II of Act 129, several EDCs provided "kits" to customers in their low-income programs. The staff 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 4 CFLs, a furnace whistle, and an LED nightlight, this should be counted as three measures (CFL, furnace whistle, and LED nightlight) when calculating the proportion of measures offered to the low-income sector.

EDCs should use the foregoing information as guidance for examining compliance with regard to 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 III. If an EDC's 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 2016 TRC 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.[10] The PUC established the TRC Order to provide guidance, methodology, and formulas for properly 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 TRC Test for Phase III will be applicable throughout Phase III, unless the PUC determines a need to modify the TRC during Phase III.

2.2.2 2016 TRC Order

Although much of the 2016 Phase III TRC Test Order (issued June 11, 2015) is consistent with the Phase II TRC Order, there are some refinements and additional guidelines. Updates and refinements to the 2016 Phase III TRC Test include the following:

- Inclusion of all reasonably quantifiable savings associated with water and fossil fuel avoided costs
- T&D avoided costs will be based on the calculation approach and data used by the SWE in its 2015 DR Potential Study.[11] Furthermore, EDCs shall use the Base Residual Auction (BRA) capacity price for a given delivery year.
- The peak demand reductions achieved by demand response programs in Phase III must be monetized by EDCs for purposes of the TRC Test.
- Guidance on an escalation factor for natural gas prices as well as the basis for the discount rate used in TRC calculations
- Adoption of the 75% participant cost assumption set forth in California's 2010 DR Cost-Effectiveness Protocols. Under this protocol, 75% of the customer incentive payment will be used as a proxy for the participant cost when calculating the TRC Test ratio for demand response programs. For EDCs that elect to use CSPs to implement DR programs when the exact incentive payment from the CSP to the participant is unknown, the EDCs are permitted to use 75% of the payment amount to the CSPs as a cost in the TRC Test.

[10] 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.

[11] Act 129 Statewide Evaluator Demand Response Potential for Pennsylvania - Final Report – Dated February 25, 2015. Released via Secretarial Letter, at Docket No. M-2014-2424864, on February 27, 2015. The report is available at Act 129 Statewide Evaluator (SWE) website: http://www.puc.pa.gov/filing_resources/issues_laws_regulations/act_129_information/act_129_statewide_evaluator_swe_.aspx

The 2016 TRC Test 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 III will be determined using gross verified savings.

All EDCs' EE&C plans are required to include both net[12] and gross TRC ratios at the program level separately for EE and DR goals. While the Commission will continue applying the TRC Test at the plan level, the Commission will continue to reserve the right to reject any program with a low TRC test ratio.

2.2.3 Incremental Costs

The Final Order for the TRC Test for Phase III of Act 129 EE&C programs ruled that incremental measure costs data will be defined for Phase III as they were for Phase II. EDCs have the flexibility to choose between the values in the SWE incremental costs database, adjusted values from the DEER database, or the values currently used for program planning and cost-effectiveness testing.[13]

2.2.4 TRC Order Schedule

The PUC issued a Final Order for the TRC Test for Phase III of Act 129 EE&C programs on June 11, 2015, and determined that the 2016 TRC Test shall apply for the entirety of Phase III. 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 in order to compare the actual Phase III benefits and costs to the planned Phase III benefits and costs using a definition of TRC costs and benefits that remains constant over Phase III.

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 Technical Reference Manual (TRM) for the State of Pennsylvania on October 3, 2005.[14] The PUC also directed the Bureau of Conservation, Economics, and Energy Planning (CEEP)[15] to oversee the implementation, maintenance, and periodic updating of the TRM.[16] On January 16, 2009, in the Energy Efficiency and Conservation Program Implementation Order for Phase I of Act 129's EE&C Program,[17] the PUC adopted the TRM as a component of the EE&C Program evaluation process. In the Phase I Implementation Order, the PUC also noted that, "as the

[12] The PUC's Phase III Implementation Order required the inclusion of net TRC ratios, in addition to gross. EDCs were to include language clarifying the speculative nature of NTG estimates. See Phase III Implementation Order at page 107.
[13] The incremental cost database is posted to the SWE Team SharePoint site.
[14] Order entered on October 3, 2005, at Docket No. M-00051865 (October 3, 2005 Order).
[15] As of August 11, 2011, the Bureau of CEEP 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.
[16] See October 3, 2005 Order at page 13.
[17] See Energy Efficiency and Conservation Program Implementation Order at Docket No. M-2008-2069887, (Phase I Implementation Order), at page 13, entered January 16, 2009.

TRM was initially created to fulfill requirements of the AEPS Act, it will need to be updated and expanded to fulfill the requirements of the EE&C provisions of Act 129."[18] Soon after the adoption of the EE&C Program Phase I Implementation Order, PUC staff initiated a collaborative process to review and update the TRM with the purpose of supporting both the AEPS Act and the Act 129 EE&C Program that culminated in the adoption of the 2009 TRM at the May 28, 2009 public meeting.[19] In adopting the 2009 TRM, the PUC recognized the importance of updating the TRM annually.[20] A program evaluation group (PEG)[21] was formed to, among other things, provide guidance to the SWE in clarifying energy savings measurement protocols and plans by recommending improvements to the existing TRM and other aspects of the EE&C program. In addition, the PUC convened a Technical Working Group (TWG)[22] meeting to discuss the proposed TRM updates.[23] In the Phase III Final Implementation Order, the PUC stated that the 2016 TRM is applicable for the entirety of Phase III. The PUC, however, reserved the right to implement a mid-phase TRM update if deemed necessary. The PUC expressed a belief that the manual has reached a level of stability whereby it provides accurate measurements of reductions. It noted in the Phase II Implementation Order that the TRM should reflect the "truest savings values possible" and should "ensure that Act 129 monies are being spent to acquire real energy savings, not fictitious savings values that only serve to protect the EDCs from potential penalties."

During Phase II of Act 129, the PUC filed and approved the 2013, 2014, 2015, and 2016 TRM Orders. The 2016 TRM Order is effective June 1, 2016, and applies to the remainder of Phase III. The approval date of the 2016 TRM is June 8, 2015, and the effective date is June 1, 2016 – May 31, 2021. Previous TRM orders and TRM manuals can be accessed through the PUC website.[24]

The approval date of the TRM is when the TRM Order was entered after the PUC approved it during a public meeting; this differs from the effective date of the TRM, which specifies when the TRM shall be used.

For Phase III of the Act 129 EE&C program, the PUC again adopted the TRM as a component of the EE&C Program evaluation process. 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

---

[18] Ibid.
[19] See Implementation of the Alternative Energy Portfolio Standards Act of 2004: Standards for the Participation of Demand Side Management Resources – Technical Reference Manual Update Order, at Docket No. M-00051865, (2009 TRM), entered June 1, 2009.
[20] Ibid., pages 17 and 18.
[21] The PEG is chaired by PUC staff and comprises representatives from the EDCs and the SWE to encourage discussions of EDC program-specific issues and associated evaluation, measurement, and verification.
[22] The TWG is chaired by PUC staff and comprises representatives from the EDCs, the SWE, and other interested parties to encourage discussions of the technical issues related to the EM&V of savings programs to be implemented pursuant to Act 129.
[23] The PUC held TWG meetings to provide stakeholders with the opportunity to review proposed high-impact changes to residential, commercial, and industrial measures, and also allow for a question and answer session about those changes. Additionally, stakeholders had the opportunity to propose any other changes to the TRM.
[24] See link: http://www.puc.pa.gov/filing_resources/issues_laws_regulations/act_129_information/technical_reference_manual.aspx

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 TRM is organized into several chapters. The first chapter provides guidance and overarching rules regarding use of the TRM. The second chapter contains TRM protocols, or measure-specific methodologies for estimating energy and demand savings, for residential measures. The third chapter contains TRM protocols for commercial and industrial measures. The fourth chapter contains TRM protocols for agricultural measures, and the fifth chapter addresses demand response. The TRM also contains appendices to present information that does not easily fit the template of a TRM protocol.

2.3.1 Purposes of the TRM

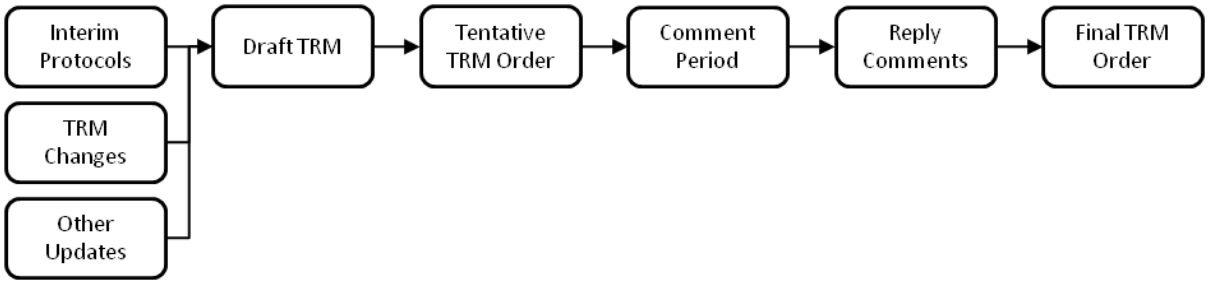
The TRM serves a variety of purposes for Act 129. In addition to providing measure savings protocols, the TRM ultimately seeks to facilitate the implementation and evaluation of Act 129 programs**.** The TRM fulfills the following objectives:

- Serves as a common reference document for energy efficiency measures to be used by EDCs, ICSPs, evaluation contractors, 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 evaluation contractors 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

For the TRM to be an effective tool for Act 129, the PUC ordered a regular annual update to the TRM in Phase I and Phase II of Act 129. 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. 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, and Final Order procedure (see Figure 2).

Figure 2: TRM Update Process



The PEG—comprising TUS staff, the SWE, EDCs, and EDC evaluation contractors—has been initiated to review, clarify, improve, and add new savings protocols to the TRM. Generally, the mission of this group is to provide technical guidance to the PUC regarding the quantification of energy and demand savings. 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. The PEG 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 PEG review process.

2.3.3 TRM Protocols

A TRM protocol 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. Measure-specific Evaluation Protocols (MEPs) have been developed to estimate energy and demand savings associated with behavioral modification and demand response programs. These MEPs are included in Section 6 of this Framework.

The Pennsylvania TRM categorizes all measures into three categories: *deemed measures, partially deemed measures,* and *custom 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.
- *Partially deemed measures* are determined using an algorithm with stipulated and open variables, thereby requiring data collection of certain parameters to calculate the energy and demand 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.

The TRM contains many protocols with deemed savings. 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 "give-away" programs.

Recommendations of the SWE 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 reviews of 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 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-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. Some open variables may have a default value to use when the open variable cannot be measured.

Open variables include the following:

- Capacity of an A/C unit
- Change in connected load
- Square footage of insulation
- Hours of operation of a facility or of a specific electric end-use

- Horsepower of a fan or pump motor

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

- Identifying high-impact measure protocols for review and providing necessary clarifications or modifications through the TRM working group based on evaluation findings, statewide studies, changes to federal and state energy efficiency codes, or more recent and reliable secondary research available.
- Analyzing algorithms and definitions of terms during the TRM update process to verify that the protocols use accepted industry standards and reasonably estimate savings.
- Analyzing low-impact measures with unrealistic and inaccurate savings values. Reviewing low-impact measures periodically to adjust the level of EM&V rigor based on market adoption.
- Ensuring that the methodologies for implementing protocols are clearly defined and can be implemented practically and effectively.
- For nonresidential measures, establishing energy impact thresholds by measure type 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 high-impact measure installations.
- Conducting Pennsylvania-specific research studies to update key assumptions for high-impact measures and provide load shapes for each measure variant.
- Examining the literature referenced in the TRM supporting key variables used in partially deemed savings algorithms which warrant further review and discussion by the PEG; this may include reviewing the population from which source data were derived, if available, and providing recommendations regarding the appropriate population or technologies to which the generalizations should be applied.

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 evaluation contractors 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.[25]

For replacements and retrofits, ICSPs will use the applicable date to determine which TRM version to select to estimate EDC claimed savings.[26] The "in-service date" (ISD) 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 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 ISD. 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 and are fully deemed. These values may not be changed or revised by ICSPs. Open variables[27] 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 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 annual reports.

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

[25] 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.
[26] 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.
[27] Open variables are listed with a default value and an option for "EDC Data Gathering" in the TRM.

vice versa. The criteria for making such changes may include the feasibility of attaining such information, the percent change in savings expected when using open versus stipulated variables, and the uncertainty surrounding default values.

For certain nonresidential end-use categories, the TRM defines thresholds where M&V is required if the threshold is exceeded. In other words, if the combined savings for a certain end-use category in a single project is 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 falls 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 evaluation contractor for determining when they must use customer-specific data gathering in order 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.

If an EDC, ICSP, or evaluation contractor believes the information in the TRM regarding a deemed or partially deemed measure should be revised, they should submit a written request to the PEG for review and consideration in the next TRM update.

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 approach described in Section 2.3.5.

2.3.4.2 Using the TRM to Determine Ex Post Savings

Typically, EDC evaluation contractors conduct research studies, site inspections, and documentation reviews based 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
TRM deemed savings measures	Follow deemed savings per TRM	Furnace whistle
TRM partially deemed measures	Follow TRM savings algorithms, using deemed variables and verified open variables	C&I lighting, residential lighting (CFLs & LEDs), C&I motor
Custom measures	Follow MEP (Section 6), applicable Uniform Methods Project (UMP) protocol or other custom measure protocol 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 correct use 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 evaluation contractor 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[28] 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

[28] Open variables are signified by the term “EDC data gathering” in the TRM.

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 nonresidential projects above a threshold kWh/yr.[29] Evaluation contractors should verify the customer-specific data for all measures in sampled projects above the threshold. If the evaluation contractor determines that the customer-specific data gathered by the ICSP are not reasonably valid, then the evaluator should conduct independent customer-specific data gathering activities for those measures. An SSMVP is required for all projects with combined measure savings above the TRM thresholds.

Section 3.3.2.3 provides additional information on nonresidential 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 evaluation contractor 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) also calculate the savings using TRM protocols and 2) include both sets of results in the EDC reports. The EDCs must explain the custom methods in the annual reports, wherein they report the deviations. If an EDC uses a custom method to calculate savings for a TRM measure, the SWE will perform a pre-approval review only 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

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, in order 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.

[29] The threshold kWh/yr is stipulated in the TRM and will vary depending on the type of measure.

2.3.5.1 Interim Protocol Approval Process

The IMP approval process is informal and 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 in the PEG. 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 PEG, the SWE gives the protocol interim approval as an "interim approved TRM 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.

The effective date of IMPs depends on the nature of the protocol. Two types of protocols have been identified: *new measure interim protocols* and *TRM modification interim protocols*. The SWE determines the appropriate classification of each proposed protocol and announces when the protocol is approved and effective.

2.3.5.1.1 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 entire program year 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.

2.3.5.1.2 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 as well as 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 EDCs, ICSP, evaluation contractor, 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/yr level.[30] 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, EDC evaluation contractors 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 to improve the EDCs' ability to support energy services that meet the EDCs' energy savings goals. EDC evaluation contractors are permitted to determine the appropriate M&V protocols for each project. EDC evaluation contractors must verify impacts for custom measures selected in the verification sample. They must develop an appropriate Site-Specific Measurement and Verification Plan (SSMVP) for each sampled project, per their professional judgment. SSMVPs should be uploaded to the SWE Team SharePoint site two weeks before the on-site inspection is scheduled by the EDC evaluator. EDC evaluation contractors 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 evaluation contractor determines that data collected by the ICSPs are not reasonably valid, 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/yr level. The evaluation contractor 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 evaluation contractors to ensure that baseline assessments are being conducted in an acceptable manner and that all necessary data points are being collected for the estimation of savings.

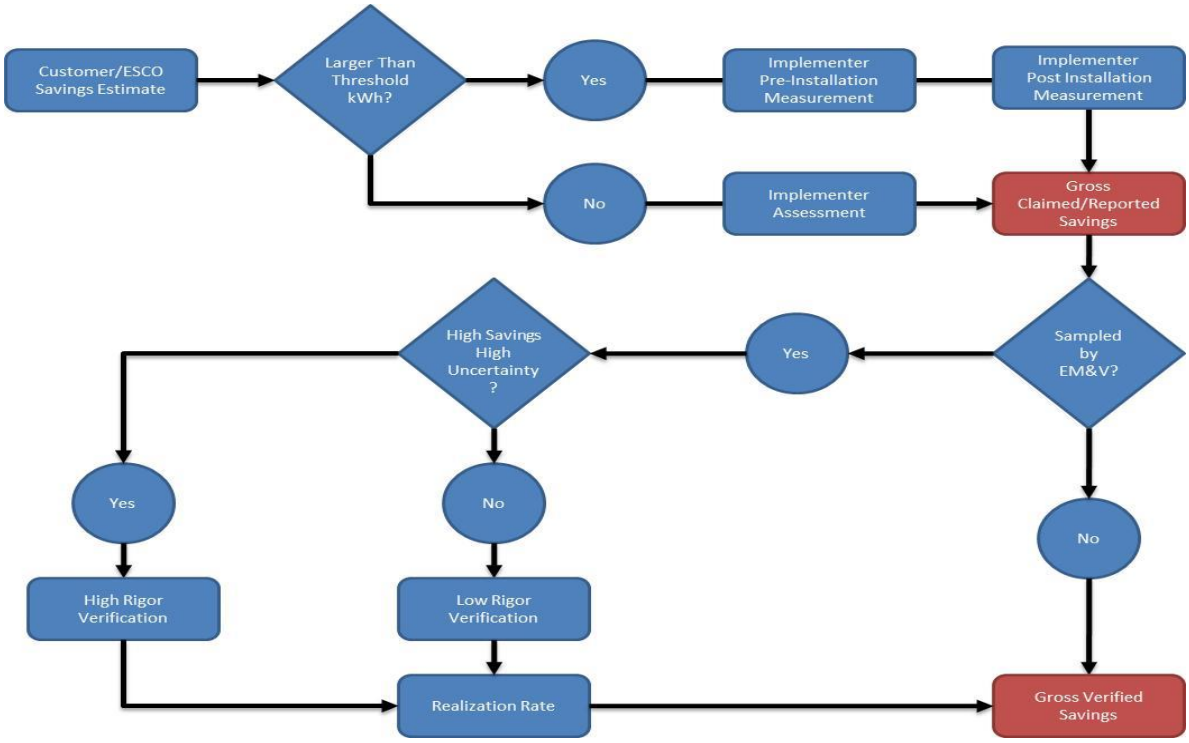
The SWE reserves the right to audit and review claimed and verified impacts of any custom measures or projects. The SWE will randomly choose projects sampled by the EDC evaluation contactors and will audit the evaluators' engineering analysis and realization rates. In addition, the SWE also may select a random sample of projects not sampled by the EDC evaluation contractors and conduct an independent assessment of the ex post savings. The SWE may use these independent samples to augment the sample selected by

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

the EDC evaluation contractors. The results from SWE independent assessments may be included in the program's realization rate calculations at the discretion of the EDC evaluation contractor.

Figure 3 presents a flow chart of the generic process to verify savings for custom measures. Deviations from the process are acceptable.[31]

Figure 3: Custom Measure Process Flow Chart



2.4 GUIDANCE MEMOS

This Evaluation Framework is developed 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 and their evaluation contractors and the 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 in the Phase III folder on the SWE Team SharePoint site. All of the guidance memos issued by the SWE in Phase II have been incorporated into this Evaluation Framework. Neither

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

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. The SWE will collaborate with EDCs primarily through collection of data from previous implementation and evaluation activities. TUS staff is responsible for approval of study memos. Results from these studies are intended to inform updates of the TRM.

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.

Section 3 Technical Guidance on Evaluation, Measurement, and Verification (EM&V)

This section of the Evaluation Framework is intended to help guide EDC evaluation contractors 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. In order to develop these estimates for a program, an EDC's evaluation contractor selects a sample of projects from the program population for verification of the ex ante savings estimate, which may include more rigorous measurement and verification activities than those used to prepare the reported savings estimates. These measurement and verification activities are discussed in Section 3.3.

A sample typically is 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 3.6.5 describes other sources of uncertainty in an evaluation and how evaluation contractors 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, high-level evaluation plan for each program and submit it to the SWE Team SharePoint site for review within 120 days of the program year's start date (by September 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 November 15 of each program year. Within four weeks of this submission, 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 9-12, or whenever intra-year changes are required. Evaluation plan updates will undergo the same review process as the initial evaluation plan for a phase of the Act 129 programs. Evaluation contractors are encouraged to submit evaluation plan modifications to the SWE as early as possible in the program year.

Each EDC and its evaluation contractor 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 also should 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 evaluation contractor, or some other entity). Evaluations plans should also explain how the EDCs would make site inspections results available to the SWE Team. Sections 3.3 through 3.7 provide technical guidance to the EDC evaluation contractors regarding evaluation plans and activities for Phase III of Act 129.

The PA TRM provides EDCs with open variables for a number of energy conservation measure (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 evaluation contractor intends to utilize the option of "EDC data gathering." 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 evaluation contractors 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 program implementers capable of reporting ex ante savings quarterly. 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. Records stored in EDC tracking systems also should be the basis of the evaluation contractor'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 annual reports.[32] 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.[33]

3.2.2 Installed Dates, In-Service Dates, Recorded Dates, Reported 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; this may or may not coincide with the In-Service Date.
- **In-Service Date (ISD, also referred to as the "Commercial Date of Operation" or CDO):** The date the measure is installed and commercially operating as intended for long term savings. This is the date at which savings begin to be realized by the customer and may be the same as the Installed Date or later. For upstream rebate programs such as lighting or appliance programs, for purposes of data tracking it is appropriate to use the transaction date as the ISD as the actual installation date is unknown.
- **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 annual report.
- **Reported Date**: The date on which savings for a given project are officially submitted to the PUC as part of an annual compliance report. The gross reported and gross verified savings values for a program quarter or program year are the sum of the measures with a Reported Date within the quarter or program year; this does not refer to the submission date of a quarterly or 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 annual report to the PUC as part of a compliance requirement.

In Phase I, an issue was identified related to reporting energy savings and more specifically, *reporting lags*. *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 PY8 but not be recorded or reported until PY9. There are two types of reporting lags: participant lag and approval lag.

[32] Cumulative savings for a time period, especially Cumulative Program Inception to Date (CPITD), may not exactly equal the sum of transactions, quarters, or program years due to adjustments to transactions, and other factors.
[33] Some worksheets used in the calculation of individual customer impacts will not be embedded in the tracking system but can be furnished upon request.

- *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.[34]
- *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 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 preliminary annual reports on July 15 and final annual reports on November 15 following the end of the program year[35] using the existing reporting structure, which accounts and works well for all projects with reported dates (and therefore in-service dates) prior to the statutory target date. EDCs opting to account for lagged transactions that have a recorded date after the statutory target date, but an in-service date prior to the statutory target date, 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 in-service date.[36] EDCs should include another table representing kW savings.

The Commission's decision to forego annual updates in favor of a fixed TRM for Phase III considerably simplifies decisions about which TRM governs savings calculations for a given project. However, situations may still arise in which it is unclear what is the appropriate TRM to use. The SWE and TUS staff agreed that the applicable date for determining which TRM to use (for all measures, excluding new construction) is the in-service date. The TUS staff and the SWE concluded that the in-service 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 evaluation contractors should use the TRM in effect at the in-service date when calculating energy and demand savings for Phase III. 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 ISD. This requirement is to account for the long lifecycle of new construction projects that are designed to a particular standard prior to construction.

---

[34] 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.

[35] *Phase III Implementation Order*, pp. 100-101

[36] Lagged transactions technically are part of later reporting periods, and therefore should not be portrayed as part of current reporting periods by including them in the actual reports.

3.2.3 Historic Adjustments

EDCs are required to document any adjustments made to ex ante savings after a semi-annual or 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 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:

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 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 evaluation contractor's responsibility to independently verify each parameter in the savings calculation. This requirement will improve the transparency of the savings calculation process. For example, to calculate energy and demand savings for residential central air-conditioning equipment using the 2016 Pennsylvania TRM, the ICSP must provide the following fields:

- Cooling capacities (output in Btuh) of the central air conditioner installed

- Energy Efficiency Ratio (EER) and Seasonal Energy Efficiency Ratio (SEER) of the qualifying unit being installed
- Energy Efficiency Ratio (EER) and Seasonal Energy Efficiency Ratio (SEER) of the baseline unit[37]
- Location of the home so that the default Equivalent Full Load Hours of operation during the cooling season can be incorporated into the savings calculation

3.2.4.1 Key Data Collection Fields for Site Visits

Audit reports provide 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. It is important to note that poor record keeping by implementation contractors has hindered analysis of savings in the past for low-income audit and weatherization programs.[38] The EDCs are encouraged to follow the 2016 LIURP codebook to the extent possible when developing a standard form for their low-income audit reports.[39] 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 CFLs or 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 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. For example:
    - Existing lamp and replacement CFL or LED wattage, and room where the CFL or LED is installed
    - Existing and replacement air conditioner capacity, model number, efficiencies, etc.
    - Existing and replacement refrigerator type, model number, wattage, etc.
    - Number of faucet aerators and showerheads

[37] This assumes that an “early replacement” savings protocol is followed.
[38] Statewide Evaluation (SWE) Team. 2014. Quantitative Comparison of Low-Income Weatherization Contractor Performance. Submitted to the PA PUC, July 21, 2014.
[39] The Pennsylvania Public Utility Commission, Bureau of Consumer Services. 2015. LIURP Codebook for the Low Income Usage Reduction Program. The codebook is posted to the SWE Team SharePoint site.

- 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 evaluation contractors that conduct gross impact evaluations. Impact evaluations determine program-specific benefits, which include reductions in electric energy usage, electric demand, and avoided air emissions[40] that can be attributed directly to an energy efficiency program. 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.

For C&I programs, impact evaluation contractors use 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 evaluation contractor(s) must collect the appropriate data for sampled projects. The EM&V activities may include surveys 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 confirming the claimed savings.

For residential programs, approved impact evaluation methods for the Act 129 residential-sector programs have evolved over the course of the Pennsylvania Act 129 programs. The Act 129 residential programs are mostly mass market programs that involve proven and well-tested technologies marketed to most or all households in a service area. As a result, ex ante estimates of gross program savings usually can be calculated using algorithms listed in the applicable Pennsylvania TRM, Interim Measure Protocols (IMP), Mass Market Protocols (Section 6), or a combination of the above for whole-house, comprehensive programs. Basic levels of rigor are typically applied when verifying residential measures. EDC implementation contractors or EDC evaluators then conduct inspections or desk

[40] While EDCs are not required to report air emissions in EE&C program impact evaluations, estimates of emission reductions can be estimated easily, based on verified gross energy savings and emissions factors from sources, such as PJM, the Energy Information Administration, and the Federal Energy Regulatory Commission.

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 results of the verification activity.

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 evaluation contractors helps reduce or eliminate redundant data collection efforts when appropriate. The 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. Savings estimation for mass market programs or non-TRM measures should follow the protocols in this framework (Section 6) or custom measure protocols developed by the EDCs.

3.3.2 Calculating Verified Gross Savings

One of the primary research objectives of an impact evaluation is to calculate gross verified savings, which are the savings achieved by the program as calculated by an independent third-party evaluator. Evaluation contractors should produce an independent estimate of program 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*. Evaluators should verify that an appropriate level of measurement rigor was employed by the ICSP, and if needed, conduct independent end-use level measurements for high-impact and high-uncertainty projects. 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 also will 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 anticipated by the Act 129 programs should be estimated 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) dictates 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 actually 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.

A large portion of these savings are either: 1) deemed based on units installed, sold, or given away, or 2) partially deemed and subject to assumptions relative to the performance of the technologies and how the technologies are used. Though metering studies and detailed analysis are encouraged to inform updates of TRM savings protocols, EDC evaluation contractors must verify fully deemed measures with TRM protocols by using TRM protocols and assumptions. Metering, building energy simulations, or other project-specific data collection activities may be required for partially deemed measures with greater variance in end-use operating parameters and custom measures.

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 evaluation contractors 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 evaluation contractors 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.

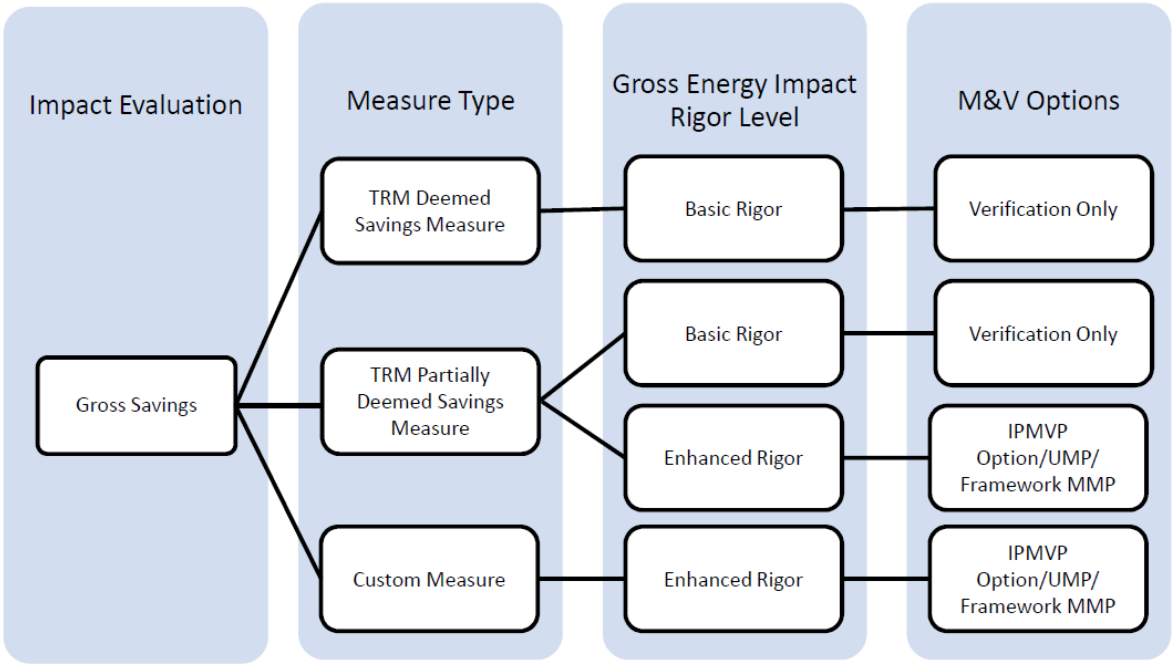
The Evaluation Framework defines two levels of rigor: *basic* and *enhanced*. Each level of rigor provides a class of minimum allowable EM&V methods, based on standard evaluation practices, in order to offer flexibility for the evaluation contractors to assess and propose the most accurate and cost-effective 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, evaluation contractors are allowed to choose the appropriate level of rigor, as long as they follow the guidelines in this section, 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 evaluation contractors and request revision of the verification technique prior to the evaluators' site visit, if necessary. After the site visit, 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 independent evaluators 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.[41] The TRM 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). These paths are depicted in Figure 4, which provides guidance on choosing the level of rigor by measure type.

Figure 4: Expected Protocols for Impact Evaluations



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

Table 14: Required Protocols for Impact Evaluations

Rigor Level	Minimum Allowable Methods for Gross Impact Evaluation
Basic	1. 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. 2. Verification of appropriate application of the TRM savings algorithms for TRM partially deemed measures using gathered site data that typically is limited to performance specification data and does not need to be measured onsite.
Enhanced	1. Simple 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 evaluation contractor, or obtained by the evaluation contractor directly. 2. Retrofit Isolation Engineering methods as described in IPMVP Option B. 3. A regression analysis (IPMVP Option C)[42] of consumption information from utility bills with adjustments for weather and overall time period reported. The SWE Team recommends that at least twelve (12) 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 residential new construction. In these cases, well-matched control groups and post-retrofit consumption analysis are allowable. 4. Building energy simulation models as described in IPMVP Option D. 5. MEP defined in Section 6 of the Evaluation Framework

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 evaluation contractors should be aligned to increase the correlation of ex ante and ex post savings estimates to improve the precision of evaluation results. Evaluation contractors can leverage information collected by the program ICSPs in cases where it would be burdensome to the participant for the

[42] 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.

evaluation contractor 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.

3.3.2.2.1 Energy – 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 actually are installed and operational, and 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, or via a review of project documentation. For each program, EDC evaluation plans should specify whether onsite inspections are planned, and if so, whether evaluation contractors 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 3.6.4.

Energy efficiency kits require special attention because installation rates have been found to be relatively low.[43] EDC evaluation contractors should independently verify the installation rate of kit measures by sampling kit participants. Stratification by measure or kit 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. While incorporating installation rates, measure savings will be calculated based on TRM values.

3.3.2.2.2 Energy – Basic Rigor Option 2: Simple 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 site data without onsite measurement. If the ICSP collects the project-specific information, evaluation contractors should attempt to confirm the accuracy and appropriateness of the values. This option should be used for partially deemed measures producing savings above the threshold

[43] *Pennsylvania Power Company Program Year 6 Annual Report,* November 2015. https://www.firstenergycorp.com/content/dam/customer/Customer%20Choice/Files/PA/tariffs/PP-PY6-Report.pdf

values[44] identified in the TRM as requiring customer-specific data collection, but which have low uncertainty.

3.3.2.2.3 Energy – Enhanced Rigor Option 1: Simple Engineering Model With Measurement

The first class of allowable methods for enhanced rigor is a Simple Engineering Model (SEM) with measurement of key parameters. An SEM 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. SEMs are straightforward algorithms for calculating energy impacts for measures such as energy-efficient 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.

3.3.2.2.4 Energy – 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 parameters for the energy use 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 lighting retrofit where both power draw and hours of operation are logged.

3.3.2.2.5 Energy – 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: 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. The EDC evaluation contractors are free to select the most appropriate additional variables to include. In certain cases, selecting matching control groups may be required to calculate differences between the treatment (participant) and control groups' pre and post consumption. A control group comparison approach is beneficial to isolate non-programmatic, extraneous effects and determine the true impact of the program intervention. The EDC evaluation contractors are required to adhere to the guidelines and protocols in Section 3.3 of this Evaluation Framework.

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

[44] Thresholds will apply only to nonresidential measures.

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 Uniform Methods Project (UMP) Protocols, which are consistent with the IPMVP Option C – Whole Facility for annual energy savings and coincident peak demand savings, respectively.[45] The UMP recommends utilizing a billing analysis to estimate total savings when multiple measures and retrofits have been installed on site in order to capture the combined effects of the installed measures or when the measure is anticipated to yield substantial savings.

3.3.2.2.6 Energy – 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 heating, ventilation, or air conditioning (HVAC) end use. This method often is 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.

3.3.2.2.7 Demand – Basic Rigor

The basic rigor level for the gross demand impact protocol prescribes that, at a minimum, on-peak demand savings be estimated based on the allocation of gross energy savings through the use of allocation factors, coincidence factors, or end-use load shapes during the hours of 2:00 p.m. to 6:00 p.m. on non-holiday weekdays (from June 1-August 31). For TRM deemed measures, TRM deemed coincidence factors are to be used. The use of TRM deemed coincidence factors should be applicable only to the TRM deemed and partially deemed measures that meet the requirements for basic rigor in Table 15. Custom measures should follow an enhanced rigor approach. Demand Response programs should follow the Demand Response M&V Protocol in Section 6.2..

The SWE encourages EDC evaluation contractors to recommend improved coincidence factors values using a load shape from metered or vetted sources, when applicable, during TRM working group discussions. The SWE will consider the proposed values for

[45] *International Performance Measurement & Verification Protocol (IPMVP); Concepts and Options for Determining Energy and Water Savings: Volume 1.* Prepared by Efficiency Valuation Organization, www.evo-world.org. September 2009. EVO 10000 – 1:2009. and Uniform Methods Protocols: Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol. http://energy.gov/sites/prod/files/2013/11/f5/53827-8.pdf

prospective TRM updates. The SWE reserves the right to request additional documentation to investigate the applicability of the load shapes submitted.

3.3.2.2.8 Demand – Enhanced Rigor

The enhanced rigor level for the gross demand impact protocol requires primary data from the program participants. These data could be interval-metered data, either from TOU consumption billing data (if appropriate), 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.

For energy efficiency measures that produce savings during peak periods, end-use interval meter data, if available, 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. If end-use interval meter data are not available, spot metering/measurement at peak pre- and post-retrofit should be conducted to assess impacts during non-holiday weekday afternoons from 2:00 p.m. to 6:00 p.m. during summer months (June 1-August 31). These data will be used with one of two engineering modeling approaches: 1) full measurement IPMVP Option B or 2) calibrated engineering model Option D, where the modeling approach must meet all requirements in the IPMVP protocol. Demand Response programs should follow the Demand Response M&V Protocol in Section 6.2.

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 evaluation contractor. The actual number of strata used will be at the evaluation contractor'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, below), 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 evaluation contractors 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 Nonresidential Programs[46]

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 Value of Information. As a guide, enhanced rigor should be used if the measure meets one or more of the following criteria: 1. Irregularity of loads: a pattern does not exist sufficient enough to predict loads with ease and accuracy 2. Irregularity of operating periods: a pattern does not exist sufficient enough 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) 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 taken into account in the TRM, exist between measures. An interactive effect is considered significant if the EDC evaluation contractor 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 have a 50/50 mix of basic and enhanced levels of rigor.
Stratum 3 – Low-Impact Measures	Basic rigor

[46] Certain mass market programs, like behavior modification, should follow the protocols in Section 6.

* The EDC and evaluation contractor may determine the appropriate level of impact and uncertainty when stratifying measures. The EDC and evaluation contractor'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 Basic Rigor EM&V Activities

3.3.3.1.1 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 EDC evaluator should verify that TRM assumptions are appropriate for the measure delivery option being evaluated.

3.3.3.1.2 Measure Installation Verification

The objectives of measure installation verification are to confirm that the measures actually were installed, the installation meets reasonable quality standards, and the measures are operating correctly and have the potential to generate the predicted savings during compliance years. At a basic level of rigor, phone interviews, combined with appropriate invoices and manufacturer specification sheets, may be used to verify the measure type.

During Phase III of Act 129, measure installation verification will follow the methodology set forth in the Market Potential Study conducted for Phase III. According to that methodology, if the evaluation contractor finds that a measure was uninstalled or not currently operating, but the ICSP reported that the measure was installed and correctly operating, appropriate savings shall still be allotted to the measure. In future Phases of Act 129, measure installation verification will continue to follow the methodology used in the corresponding Market Potential Study.

If the evaluation contractor 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 evaluation contractor 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.

3.3.3.2 Enhanced Rigor EM&V Activities

3.3.3.2.1 Baseline Assessment

Where applicable and appropriate, it will be recommended to conduct pre-installation inspections to verify the existing equipment and gather the equipment baseline data in order 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 existing equipment values should be verified by onsite inspection when possible, and replace-on-burnout existing equipment values should be based on local or federal minimum codes and standards.

3.3.3.2.2 Measure Installation Verification

As discussed in the basic rigor EM&V section, the objectives of measure installation verification are to confirm that the measures actually were installed, are operating correctly, and have the potential to generate the predicted savings during compliance years. Similarly, measure installation verification will follow the methodology set forth in the Market Potential Study conducted for Phase III. According to that methodology, if the evaluation contractor finds that a measure was uninstalled or not operating, but the ICSP reported that the measure was installed and correctly operating, appropriate savings shall still be allotted to the measure. In future Phases of Act 129, measure installation verification will continue to follow the methodology used in the corresponding Market Potential Study.

Evaluation plans should describe site inspections planned for residential and nonresidential programs. At an enhanced level of rigor, measure installation should be verified through onsite 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, evaluation contractors should note CFL applications in seldom-used areas or occupancy sensors in spaces with frequent occupancy during measure verification activities then modify hours-of-use categories appropriately. Further, if the evaluation contractor 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 evaluation contractor 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.

3.3.3.2.3 Onsite Sampling of Installations

This section provides guidance in determining the number of installations to verify during the onsite 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 evaluation contractors are encouraged to propose other options in their program evaluation plans.

The first method is to verify a census of all of the installations onsite. This activity is to be done in cases where a limited number of installations were made, or when the variance in operating parameters is large and impacts are high and need to be documented in combination with the verification activity of the evaluation contractor. 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 building should be grouped according to similar usage patterns, thus reducing the expected variability in the measured quantity 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 onsite installations is ± 20% at the 90% confidence level at the facility level. The sampling unit (line item on the TRM Appendix C form,[47] condensing unit, appliance, etc.) should be identified in the Site-Specific Measurement and Verification Plan (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 C_v 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 facility. For example, lighting retrofits in common areas should be separated from those in individual suites in an office building, or air handler unit (AHU, 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 onsite counting exercise, an error band[48] 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 evaluation contractor 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 evaluation contractor verifies only 210 fixtures in the facility hallways, ex post savings values should be calculated based on the evaluator's observations.

[47] http://www.puc.pa.gov/pcdocs/1370271.xlsx

[48] 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).

3.3.3.2.4 Site-Specific Measurement and Verification Plan

A Site-Specific Measurement and Verification Plan (SSMVP) is designed to specify the data collection techniques for physical evidence or survey responses from field installations of energy-efficient technologies. 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 evaluation contractors must design and document SSMVPs for each measure and define the quantitative data that must be collected from the field or other primary sources. SSMVPs are required for projects with combined measure savings above the TRM thresholds and are encouraged for all projects. The SSMVP should cover all field activities dedicated to collecting site-specific information necessary to calculate savings according to the engineering equations specified at the project level and to prepare for an evaluation audit of gross savings impacts. This procedure includes specifying data to be gathered and stored for field measurements that document the project processes and rationale. For non-custom measures, general measure-specific data collection workbooks may be used for preparing and completing onsite visits. For custom measures, the SSMVP should include a full narrative describing all of the associated evaluation activities and ensuing calculations. These activities typically include:

- 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.

EDC evaluation contractors should assess the expected uncertainty in the end-use energy consumption variables and develop an SSMVP for a sampled custom measure that manages the uncertainty in the most cost-effective manner. The contribution of specific engineering parameters to the overall uncertainty in the savings calculations should be identified and used to guide the development of the SSMVP.

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
8. Recording and Data Exchange Format

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:

- Relevant building configuration and envelope characteristics, such as building floor area, conditioned floor area, number of building floors, opaque wall area and U-value, window area, and solar heat gain coefficient;
- Relevant building occupant information, such as number of occupants, occupancy schedule, and building activities;
- Relevant internal loads, such as lighting power density, appliances, and plug and process loads;
- Type, quantity, and nominal efficiency of relevant heating and cooling systems;
- Relevant HVAC system control set points;
- Relevant changes in building occupancy or operation during the monitoring period that may affect results; and
- Description of the energy conservation measures 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 onsite 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 evaluation contractors have considerable latitude regarding the development of an SSMVP, which may be a combination of the IPMVP options.

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 and documentation supporting baseline assumptions as part of the SSMVP may be presented in the most convenient format (spreadsheet or written report), but should always be clearly stated and explained. This aspect is a key component of an SSMVP, in addition

to the application documents. Fully specifying the data analysis procedures will help ensure presentation of an efficient and comprehensive SSMVP.

Field Monitoring Data Points: If any actual field measurements are planned, they should be specified, including the sensor type, location, 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. This section may also discuss non-measured data sources. For example, in a situation where the evaluation contractors intend to 'annualize' savings using a comparison of the production levels from a plant during the M&V period to some 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.

Recording and Data Exchange Formats*:* Data formats compliant with the data reporting guidelines described in Section 4.1 of this Evaluation Framework should be specified.

3.4 NET IMPACT EVALUATION

The PUC stipulated in the Phase III Implementation Order that NTG adjustments be treated the same way for Phase III as they were during Phase I and Phase II. The Commission stated that "*NTG is used for making modifications to existing programs in the current phase, as well as for planning purposes for future phases*" and that "*EDC compliance with targets* [is determined] *through the use of gross savings*." [49]

The PUC says NTG should not be used for compliance "because net-to-gross ratios can vary significantly for a program from year-to-year and due to Commission and SWE concerns about relying on NTG research results to determine compliance and possible penalties for EDCs."[50]

The PUC, however, recognizes that net savings are valuable for informing program modifications and program planning and for determining program cost-effectiveness, and that "*the inclusion of NTG-based TRC ratios would provide all stakeholders with additional information regarding the effectiveness of EE&C measures and programs*." [51]

EDCs' evaluation contractors 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.

---

[49] Pennsylvania Public Utility Commission, Energy Efficiency and Conservation Program Implementation Order, at page 103, at Docket No. M-2014-2424864, (Phase III Implementation Order), entered June 11, 2015.
[50] Ibid., p. 105.
[51] Ibid., p. 105-106.

When conducting NTG research, the NTG methods should be consistent across time and EDCs. 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. 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 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 annual reports.

3.4.1 Acceptable Approaches to Conducting NTG Research

NTG research traditionally has two primary purposes: 1) attribution –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 NTGR.

The Uniform Methods Project (UMP) provides the following relevant definitions: [52]

- **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, 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." [53]

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

[52] 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, September 2014. http://energy.gov/sites/prod/files/2015/02/f19/UMPChapter23-estimating-net-savings_0.pdf

[53] 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 November 11, 2014 from http://www.iepec.org/conf-docs/conf-by-year/2013-Chicago/095.pdf.

tracing, many of which may be used to assess market effects. The Uniform Methods Project details these various methods.[54] Much has been written about the various methods and their relative strengths and weaknesses.[55] 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 increasingly 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 is problematic. In addition, while increased participant and nonparticipant 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.[56]

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 III the EDCs may use methods of their own choice, including market effects approaches, to estimate NTG for upstream programs. Section 3.4.1.5 presents a common set of methods for upstream lighting programs. The common upstream lighting program NTG methods allow some flexibility for individual EDCs. They include Market Progress Indicators (MPIs) to assess overall market progress, and options for NTG methods.

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, 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

[54] Ibid.
[55] A general review of issues and recent bibliography is provided in Haeri, H. and M. Sami Khawaja, "The Trouble with Freeriders," *op cit*.
[56] 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.

decision that self-report-based NTG should not be used to calculate net savings estimates for compliance purposes.[57] However, careful application of social science methods may help mitigate biases.[58] 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 in order to invest in energy efficiency. Thus, 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 in order 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 nonparticipants 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 nonparticipants who take actions the program recommends but without program support (outside or nonparticipant spillover).

The NTG ratio removes free ridership from the savings calculation and adds program spillover. The NTG formula is defined in Equation 2:

Equation 2: 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 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

[57] Ibid.
[58] Haeri, H. and M. Sami Khawaja "The Trouble with Freeriders." *Public Utilities Fortnightly*. March 2012 (http://www.fortnightly.com/fortnightly/2012/03/trouble-freeriders).

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[59] 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.

Care must be taken when developing the questions used to measure free ridership. The SWE considers the research approaches detailed in the UMP[60] as well as those used in Massachusetts[61] and by Energy Trust of Oregon[62] to constitute some of the best practices for free ridership and spillover estimation.

3.4.1.1.1 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.

In early Phase II, the SWE developed common methodologies for estimating free ridership in downstream programs that EDCs should use or adapt to their purposes. 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 chosen by 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.[63] The approach for appliance recycling programs is based on the approach described by the U.S. Department of Energy's Uniform Methods Project. Both approaches have undergone detailed review by the PEG.

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

---

[59] 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. http://uc-ciee.org/downloads/mrkt_effts_wp.pdf

[60] http://energy.gov/sites/prod/files/2015/02/f19/UMPChapter23-estimating-net-savings_0.pdf

[61] http://ma-eeac.org/wordpress/wp-content/uploads/Cross-Cutting-Net-to-Gross-Methodology-Study-for-Residential-Programs-Suggested-Approaches-Final-Report.pdf; http://ma-eeac.org/wordpress/wp-content/uploads/Massachusetts-PAs-Cross-Cutting-CI-Free-ridership-and-Spillover-Methodology-Study.pdf

[62] http://energytrust.org/library/reports/101231_Fast_Feedback_Rollout.pdf

[63] Ibid.

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 the appliance retirement program is based on the participants' self-report of what they would have done absent the program. Savings are attributed based on four scenarios: 1) they would have kept the unit in the absence of the program but instead, as a result of the program, replaced it with a more efficient one (savings equals delta energy usage from old to new unit); 2) 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); 3) 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 10 years old to a retailer (savings equals a mix of full savings, delta old to new, and no savings); or 4) in the absence of the program, they would have taken the unit out of usage, sold or given a unit at least 10 years old to a retailer, hauled it to the dump, or hired someone to discard it (free rider – no savings).

The SWE produced memos describing the common approaches, which are included as Appendix B and Appendix C of this Framework. The memos describe both the general form of questions to use and rules for calculating free ridership scores from responses to questions. As described in the memos, 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.

3.4.1.1.2 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 nonparticipant spillover is desirable, it is not required. If assessed, nonparticipant spillover may be assessed through either a general population (nonparticipant) survey or a survey of trade allies.

In early Phase II, the SWE developed a common methodology for estimating participant and (if EDCs choose to assess it) nonparticipant spillover in downstream programs. The SWE produced a memo describing the common approaches, which is included as

Appendix D. The memo describes both the general form of questions to use and rules for calculating spillover scores from responses to questions. The memo 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.

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

For EDCs that choose to assess it, nonparticipant spillover may be assessed either through a general population (nonparticipant) 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 nonparticipant spillover through trade ally surveys, separate surveys should be conducted for the residential and nonresidential 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 nonparticipant 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 nonparticipant spillover is assessed, the sampling approach should produce a sample that is representative of the target population (nonparticipants 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 nonparticipant 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 (nonparticipant) groups, such as with large-scale opt-out programs.[64] 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 onsite verification of the energy efficiency level of the equipment and a survey of both participants and nonparticipants. 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 nonparticipants 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 nonparticipant and participant groups. Often for low-income programs, an econometric model uses rolling-enrollment to capture participation effects.

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, it is not possible to disaggregate free riders or

[64] 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.

to identify spillover, so approaches using econometric modeling provide a hybrid estimate between gross and net savings and do 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."[65] Only certain programs can be expected to generate 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 are numerous; the programs target *markets* rather than program participants; the programs aim to change energy use through changing what happens among upstream market actors, rather than focusing just on end-users of equipment or services; 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.[66]

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. 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 four factors to consider in conducting market effects studies, whenever they are appropriate based on the above criteria.[67]

1. There needs to be 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, as well as development of metrics or market progress indicators (MPIs) against which the progress of the program in effecting change in the market may be assessed.

---

[65] 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.

[66] 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.

[67] Hoefgen, L., A. Li, and S. Feldman. *Asking the Tough Questions: Assessing the Transformation of Appliance Markets*. *Proceedings of the American Council for an Energy-Efficient Economy Summer Study on Buildings*. In Volume 10, pp. 14-25. August 2006. Herman, P., S. Feldman, S. Samiullah, and K. S. Mounsih. *Measuring Market Transformation: First You Need A Story… Proceedings of the International Energy Program Evaluation Conference.* pp. 3.19-326. August 1997.

2. Researchers must 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.
3. "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.
4. 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 III 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, 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 High-Impact Measures (HIMs)

During PY6, the SWE suggested that EDCs oversample measure categories (technologies) of high importance, called high-impact measures (HIM), to help program planners make decisions concerning those measures for downstream programs only.[68] The SWE proposed that for each program year,[69] 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.

[68] 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.

[69] The proposed HIM-specific assessment does not change any prior *Framework* requirement regarding what EDC's evaluators should do in the event that 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.

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 Upstream Lighting

The lighting market has been changing rapidly, stimulated by the Energy Independence and Security Act of 2007 (EISA) and the EISA update scheduled for 2020, and by the rapid technological development and falling prices of LEDs. For this reason, NTG estimation for upstream lighting is at least as important as for any other program; NTG estimation is essential for determining whether, and in what form, continued use of ratepayer funds for efficient lighting is warranted. At the same time, NTG protocols for upstream lighting are more involved than for other measures because of the difficulty in identifying the purchasers of the program-discounted bulbs, the incompleteness of market sales data, and the general lack of a single method that has proven completely effective in reliably capturing the full impacts of the upstream program design. Given the importance and difficulty of upstream lighting NTG estimation, the SWE has developed a common but flexible set of methods that rely on the preponderance of evidence approach. One component of that is to assess a common set of MPIs. MPIs will help EDCs understand the status of the market in terms of residential lighting products. Many of the MPIs (e.g., sales of efficient lighting products) can be used either to feed directly into NTG estimates or indirectly to help determine NTG—and in particular market effects—by providing insight into such metrics as efficient lighting availability and awareness.

The MPIs are a critical part of "preponderance of evidence" for understanding NTG issues, and market effects are good indicators of market transformation. During PY8, MPIs can help the EDCs understand how they can best design programs in terms of maximizing the effectiveness of incentives. MPIs also can serve the future purpose of serving as inputs for developing better and more consistent NTG estimates for upstream lighting programs in Pennsylvania. Many of the MPIs are commonly collected through activities such as general population surveys, customer intercepts, or supplier interviews, which minimizes the added cost of data collection task for NTG. Guidance on how MPIs can be used not only to track progress in the market, but also to help inform NTG estimates, is provided in an August 31, 2015, memo from the Phase II SWE.[70]

[70] Peters, Jane, Ryan Bliss, and Scott Dimetrosky, "Lighting Net-to-Gross Methods." Memo provided to the EDC Evaluation Teams, August 31, 2015.

Based on this memo, the SWE expects EDCs to report on some of these MPIs during PY8, whenever the MPI provides value to the EDC planning or evaluation—some of which might be better estimated on a statewide level than at the EDC level:[71]

- Awareness of the program and program technology (general population survey)
- Satisfaction with the program and technology (general population survey)
- Preference/intention to purchase (general population survey)
- Availability (shelf survey / supplier interviews)
- Pricing (shelf survey / supplier interviews)
- Attractiveness (general population survey)
- Willingness to pay (general population survey / intercept survey)
- Quality (general population survey / shelf survey)
- Sales (in-home survey / point-of-sale [POS] data)[72]
- Intention / likelihood to purchase in absence of the program (supplier interviews)
- Penetration (on-site survey) [73]
- Saturation (on-site survey)[74]

The Net Savings chapter of the Uniform Methods Project (UMP) says that best practices for estimating NTG involve using multiple methods.[75] This is especially important for upstream lighting programs, with potentially large but very uncertain savings. As pointed out in the Phase II SWE's "Lighting Net-to-Gross Methods" memo, all methods have their strengths and weaknesses, and no single method can be relied upon by itself. The SWE therefore recommends the EDCs to make use of one or more of the following methods, or variations of them, for estimating NTG for their upstream lighting programs—some of which, again, might be better estimated on a statewide level than at the EDC level:[76]

- Consumer self-reporting (general population / intercept surveys)
- In-depth interviews or surveys with lighting suppliers (manufacturers, high-level retail buyers, and store managers)
- Demand elasticity modeling (program price and sales data analysis / customer intercept surveys)
- Comparison area analysis (POS and consumer panel data, available from CREED[77] / comparison of saturation in different areas)

---

[71] The SWE and the TUS will discuss which MPIs are more appropriately estimated on a statewide level rather than the EDC level

[72] The memo lists a general population survey as an option for estimating this parameter, but results from this method have proven to be extremely unreliable and the Phase III SWE recommends in-home surveys or point of sale data instead of a general population survey.

[73] Penetration: percentage of households with at least one LED installed, or with at least one in storage

[74] Saturation: percentage of installed general purpose bulbs across households that are LEDs; originally labeled as penetration in the memo

[75] 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, September 2014.

[76] The SWE and the TUS will discuss which NTGRs are more appropriately estimated on a statewide level rather than the EDC level.

[77] Consortium for Retail Energy Efficiency Data (CREED): http://www.creedlighttracker.com/ . Note that a limitation listed in the Phase II SWE's "Lighting Net-to-Gross Methods" memo—the lack of data from home

- Long-term market effects modeling (saturation data and Bass diffusion curve modeling)
- Delphi panel to review NTG estimates from other methods and arrive at a final recommended NTG

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

improvement channels—has been greatly ameliorated in the latest data available from CREED by the combination of POS data with purchase data from hundreds of thousands of households participating in a national consumer panel.

evidence of progress in achieving program goals and objectives from the perspective of its various target audiences. 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.[78]

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, nonparticipants, end users, and trade allies. Such complex evaluations require resources and time to implement. The New York State Process Evaluation Protocols[79] 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

[78] The SWE reserves the right to review the process evaluation plans (the process evaluation plans are part of the overall EDC evaluation plan).

[79] Johnson Consulting Group. New York State Process Evaluation Protocols. Prepared for the New York State Research and Development Authority, the New York State Evaluation Advisory Group, and the New York Public Service Commission. January 2012. Accessed 4/10/13. http://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/766a83dce56eca35852576da006d79a7/$FILE/Proc%20Eval%20Protocols-final-1-06-2012%20revised%204-5-2013.pdf

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 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 nonparticipants
- Interviews and surveys with people using the technologies (e.g., usability studies of websites)
- Interviews and surveys with key policy-makers
- Observations of operations and field efforts, including field tests and investigative efforts
- Operational observations and field-testing, including process-related measurement and verification 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, nonparticipants, 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 in order 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 then compare and contrast with information stakeholders and participants express during interviews and surveys.

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

In addition to program staff, many other individuals are involved in a program, including policy-makers (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 in order 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/or Focus Groups with Participants and Nonparticipants

One purpose of virtually all process evaluations is to understand the customer's experience in order 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 also is 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 saving levels 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 are qualitative, the evaluation team's analysts must be systematic and careful in order 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 audiences 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 is expected.

Once the EDC conducts a process evaluation, the following will occur:

- The evaluation contractor's process evaluation methodology, findings, and recommendations for all process and market evaluations conducted during the year will be presented in the EDC final annual report (November 15).
- The SWE will follow up with the EDC staff to determine how the EDC plans to address each of the process evaluation recommendations made by the EDC's evaluation contractor.
- Through these conversations, the EDC will inform the SWE what, if any, action (accept, reject, still under consideration, etc.) they are planning to take based on the recommendations.
- The SWE will summarize the reports, recommendations, and the EDC's response to the recommendations in its annual report to the PUC (February).

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 the sampling uncertainty for an entire program, the reality is that M&V takes many resources, so sampling is needed. When a representative sample of

projects 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 as a whole 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 (C_v) for programs that use simple random sampling and an error ratio for programs that use ratio estimation. The C_v of a population is equal to the standard deviation (σ) divided by the mean (μ), as shown in Equation 3.

Equation 3: Coefficient of Variation

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

When ratio estimation is utilized, standard deviations will vary for each project in the population. The error ratio is an expression of this variability and is analogous to the C_v for simple random sampling.

Equation 4 provides the formula for estimating error ratio.[80] The σ term in Equation 4 is equal to the difference between the project-level verified savings estimate and the realization rate multiplied by the reported savings.

Equation 4: Error Ratio

$$Error\ Ratio = \frac{\sum_{i=1}^{N} \sigma_i}{\sum_{i=1}^{N} \mu_i}$$

Equation 5 shows the formula used to calculate the required sample size for an evaluation sample[81] based on the desired level of confidence and precision. Notice that the C_v term is in the numerator, so required sample size will increase as the level of variability increases.

Equation 5: 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

[80] Equation 4 is based on the methodology set forth in the California Evaluation Framework. The National Renewable Energy Laboratory's Uniform Methods Project (NREL UMP) provides a slightly different formula for the calculation of error ratio that is an acceptable alternative if evaluation contractors wish to use it.
[81] If ratio estimation is used, evaluators may replace C_v with error ratio in Equation 5.

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 evaluation contractor 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 evaluators can assume a default C_V equal to 0.5 for each sample population to determine target sample sizes. Once the C_V has been measured, evaluators may use that historical C_V in developing their sampling plans. Evaluators should estimate the C_V values for each sampled population and report the values in their annual reports so they can be used in subsequent evaluation plans.

The sample size formula shown in Equation 5 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 for the purposes of sampling. 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 5 by the FPC formula shown in Equation 6 will produce the required sample size for a finite population.

Equation 6: 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 7.

Equation 7: Application of the Finite Population Correction Factor

$$n = n_0 * fpc$$

3.6.1 Evaluation Precision Requirements

Table 16 provides minimum levels of sampling uncertainty prescribed for the Act 129 gross impact evaluations in order to balance the need for accurate savings estimates while limiting the costs of evaluation. The values in Table 16 assume a two-tailed design and specify the confidence and precision that must be met or exceeded each time a gross

impact evaluation is conducted. The values in Table 16 are also suggested for net-to-gross 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.

An estimate of gross verified energy savings with ±10% relative precision at the 90% confidence indicates that if evaluators resampled the same population repeatedly, 90% of the time the resulting confidence intervals would include the true value of the measured parameter,[82] assuming an unbiased sample. In reality, there are a number of other sources of uncertainty that are less straightforward to quantify and reduce the precision of savings estimates. These factors are discussed in Section 3.6.5, but should not be addressed by evaluators when calculating the achieved precision of a verified savings estimate.

Table 16: Minimum Annual Confidence and Precision Levels

Portfolio Segment	Confidence and Precision Level
Residential Portfolio	90/10
Nonresidential 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 evaluation contractors 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 III 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 upstream lighting. 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, a large C&I program, and a GNI program that all include prescriptive lighting rebates. Evaluation contractors 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 evaluation contractors submit a memo to the SWE for approval that outlines the definition of evaluation initiatives prior to drafting a complete EM&V plan.

Special consideration should be given to the following situations:

[82] Lohr, 2010.

1. Crosscutting initiatives that span both the residential and nonresidential sectors must[83] be evaluated separately, one for the residential sector and one for the nonresidential sector.
2. Evaluation contractors 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 government, non-profit, and institutional sectors within the non-residential portfolio should be treated similarly to low-income. Evaluation initiatives may include both GNI and non-GNI projects for sample design purposes with the calculated realization rate for the initiative applied to all projects, both GNI and non-GNI.
4. 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.
    - Prescriptive Lighting
    - Prescriptive Non-Lighting
    - Custom rebates
    - Direct installation
    - Business Energy Reports
5. The residential sector evaluation should include no fewer than four initiatives. Within the residential portfolio, a potential group of initiatives might be:
    a. Home Energy Reports
    b. Audits and weatherization / Whole-house program
    c. Upstream lighting
    d. Appliance Recycling
    e. School education and kits
    f. Rebates for efficient products
6. It often is more challenging to obtain accurate peak demand savings estimates than annual energy savings estimates, and peak demand savings estimates will exhibit a greater degree of variability between ex ante and ex post. The levels of precision established in Table 16 are required for energy savings estimates. If achieved precision values for peak demand impacts are significantly greater than the precision of energy estimates, evaluators should examine the source of the variation to determine if revisions to ex ante demand savings assumptions or ex post analysis techniques are warranted.

Evaluation contractors may use their professional judgment in the design of the sample as long as they meet the minimum precision requirements. Evaluation contractors should

[83] 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 nonresidential sector does not need to be evaluated as a standalone program.

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 annual confidence and precision targets are not met, corrective actions will be required in the current or subsequent evaluation year within the compliance period.

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.

Evaluation contractors 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 4. 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, 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. 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/yr ± 400 kWh/yr, remedial actions should be taken in the same program year or the following program year to improve the precision of the savings estimate. If the confidence interval at the end of the phase includes 0 kWh, no verified savings for the program should be claimed because the evaluator cannot ensure that the program impact is not equal to zero at the 85% confidence level.

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. **Estimation in the absence of auxiliary information** (also referred to as *mean-per-unit estimation*): 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 a large number of rebates for similar equipment types.

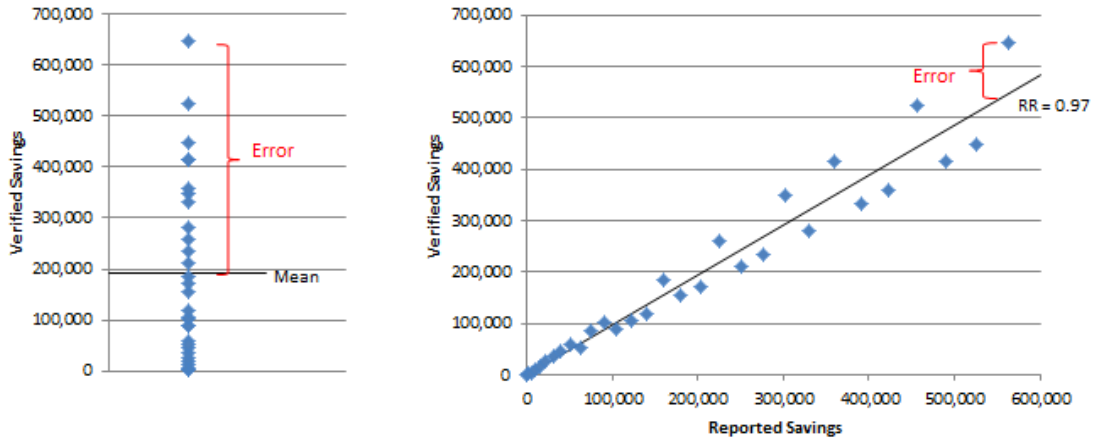
2. **Estimation using auxiliary information** (also referred to as *ratio estimation*): This is recommended for nonresidential programs, or residential programs offering a variety of measures with varying savings, because the sizes of the savings estimates of the projects within a program vary considerably 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/yr 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 the program as a whole. 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 5 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 5: 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 evaluation contractors 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 have only 10 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 evaluation contractors may find value in comparing results between strata; e.g., comparing the verification rates between measures within a program.

Evaluation contractors 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 evaluation contractor.

3.6.3 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, coefficient of variation, and relative precision. This information has been excluded from subsequent versions of the Evaluation Framework. If EDCs, their evaluation contractors, or stakeholders require additional information regarding sampling, the following resources will be helpful:

- *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.

3.6.4 Presentation of Uncertainty

There are no minimum precision requirements for EDC evaluations of Phase III savings as a whole. However, if the annual minimums established in Table 16 are met, the relative precision values of the total Phase III savings will meet or exceed the annual requirements at the same levels of confidence. In the 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 8: Error Bound of the Parameter Estimate

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

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.

$z - statistic$ = 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 no longer is 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 8 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 in order to calculate the gross verified savings for the program. The error bound of the *verified savings estimate* (in kWh/yr or kW) should be reported for each program and is calculated as follows:

Equation 9: 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[84] for each program should be reported, as well as the confidence level at which it was calculated. This formula is shown in Equation 10:

Equation 10: 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 III of Act 129, each EDC will have five verified savings estimates for energy and five verified savings estimates for demand for each program in its portfolio (one for each program and program year). The Phase III verified savings estimate is the sum of these values. These verified savings estimates will be calculated via as many as five impact evaluations. Although the error bound estimates for each impact evaluation are expressed in the unit of interest (kWh/yr or kW), they cannot be summed to produce the error bound for Phase III impacts. Equation 11 shows the formula for calculating the error bound of the Phase III impacts for a program that receives two impact evaluations—one for PY8 and PY9 and a second for PY10-PY12. 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 III 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 8 and Equation 9 using the standard error of the parameter estimate and the z-statistic associated with the 90% confidence interval (1.645).

[84] The relative precision of the verified savings estimate should equal the margin of error of the estimation parameter.

Equation 11: Phase III Error Bound

$$Error\ Bound_{Phase\ III} = \sqrt{Error\ Bound^2_{PY8,PY9} + Error\ Bound^2_{PY10-PY12}}$$

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

Equation 12: Relative Precision of Phase III Savings Estimate

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

Equation 11 also should be used to combine the Phase III error bounds from programs to the sector level and from the sector level to the portfolio level. Note that Equation 11 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.

3.6.5 Systematic Uncertainty

Section 3.6.1 of the Evaluation Framework discussed 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, largely is 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 were 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 annual reports should discuss major sources of systematic uncertainty and the efforts the evaluation contractor 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 *equivalent full load hours* (EFLH) 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. Onsite 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.

   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 Uniform Methods Project Sampling Protocols[85] (UMPSP) 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:

[85] The protocols can be found at http://energy.gov/eere/downloads/uniform-methods-project-methods-determining-energy-efficiency-savings-specific.

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?[86]
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 annual reports should discuss the steps evaluation contractors 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 is warranted.

3.7 COST-EFFECTIVENESS

Results from the EDCs' surveys and M&V activities, evaluation reports, audits, and the statewide impact evaluations will be input into a benefit/cost model and other models, as appropriate, to assess the cost-effectiveness of the EDCs' efforts at the measure, program, 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 Total Resource Cost (TRC) Test. The guidelines for the Phase III TRC are stipulated in the 2016 TRC Order. All cost-effectiveness evaluations and assessments will be conducted in accordance with the PUC's latest TRC Order.

3.7.1 TRC Method

The PUC has adopted the *California Standard Practice Manual: Economic Analysis of Demand-Side Programs and Projects* TRC Test definition, formula, and components with a few slight modifications. 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

[86] 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).

avoided monetary cost of supplying electricity is greater than the net present value of the monetary cost of energy efficiency conservation measures."[87]

According to the California manual:

> *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.*

EDC evaluation contractors should refer to the 2016 TRC Order for Phase III, and the *California Standard Practice Manual*, for detailed formulae and definitions related to the proper calculation of the PA TRC Test.[88,89]

3.7.2 Application of 15-Year Avoided Cost Streams

The TRC Order limits the effective useful life of any energy efficiency measure to 15 years for the purposes of the benefit/cost calculations but does not specifically address which 15 years of avoided costs should be used. EDCs should follow the guidelines below while developing their TRC models for Phase III of Act 129.

- The 15-year avoided cost stream for each program year should begin with the calendar year at the close of the program year using avoided costs that are calculated by calendar year. For example, for a measure installed in PY8 (June 1, 2016-May 31, 2017) with a 15-year measure life, the avoided cost stream used would be from January 2017 through December 2031.
- All EDCs should consider using short-term avoided capacity cost forecasts from the PJM Base Residual Auction for TRC calculations, since the PJM delivery year is aligned to Act 129 program years (June 1-May 31).

3.7.3 Aligning Measure Savings with Incremental Measure Costs

To determine energy efficiency cost-effectiveness using the TRC Test, the energy efficiency 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

---

[87] *California Standard Practice Manual: Economic Analysis of Demand-Side Program and Projects*: October 2001
[88] Ibid., October 2001, p. 18.
[89] Pennsylvania Public Utility Commission, *2016 Total Resource Cost Test Order*, Docket No. M-2015-2468992, June 22, 2015.

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 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 base model that otherwise would have been purchased. Similarly, the savings are calculated as the reduced energy consumption of the efficient unit compared to the base model.

Five basic measure decision types are referenced in Table 18, along with a summary of the definition of incremental measure costs and savings for each of the decision types.

Table 18: Measure Decision Types

Type of Measure	Incremental Measure Cost ($/Unit)	Impact Measurement (kWh/yr/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) minus present value of baseline device (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 2016 TRC Order defines incremental measure cost 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 Retirement calculation methodology, provided the EDC documents which method they used and why.

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 2016 TRC Order[90] or the National Action Plan for Energy Efficiency's "Understanding Cost-Effectiveness of Energy Efficiency Programs" report.[91] 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
  - Energy and peak demand line losses
  - Utility Discount Rate
  - General rate of inflation
- Program-/Measure-Specific Assumptions
  - Number of participants
  - Annual energy (kWh) and demand savings (kW)
  - Effective Useful Life
  - Incremental measure cost
  - Avoided O&M benefits (optional)
  - Outside rebates/tax credits (if quantifiable)
  - Additional direct/marketing costs[92] (non-incentive costs)
  - Program/measure load shapes
  - Measure-specific peak coincidence factor

3.8 Frequency of Evaluations

As mentioned in Section 3.5.1, every program (or initiative) should have at least one process evaluation in every funding cycle or phase; EDCs, appropriately, have not typically conducted process evaluations of every program or initiative every year. Similarly, most EDCs have not typically conducted net impact evaluations annually for every program or initiative. It has been more common to conduct annual gross impact evaluations, but the SWE believes this may not always be necessary, especially in a five-year phase. Gross impact evaluations can be staged for better use of evaluation funds in research areas of higher priority and prospective value. Initiative population from two program years can be combined into a single sample frame for initiatives that do not receive an impact evaluation every year. In such cases, a single statistically valid realization rate should be applied to the sum of reported savings for the two program years. The ex-post savings for the first of the

[90] Pennsylvania Public Utility Commission, *2016 Total Resource Cost Test Order*, Docket No. M-2015-2468992, June 22, 2015.
[91] http://www.epa.gov/cleanenergy/documents/suca/cost-effectiveness
[92] Direct or marketing costs include program administration, EDC Implementation CSP, EDC Evaluation contractor, etc.

two program years should then be reported as un-verified savings in the EDC annual report. The EDCs should use the following criteria to propose an appropriate frequency for every program or initiative:

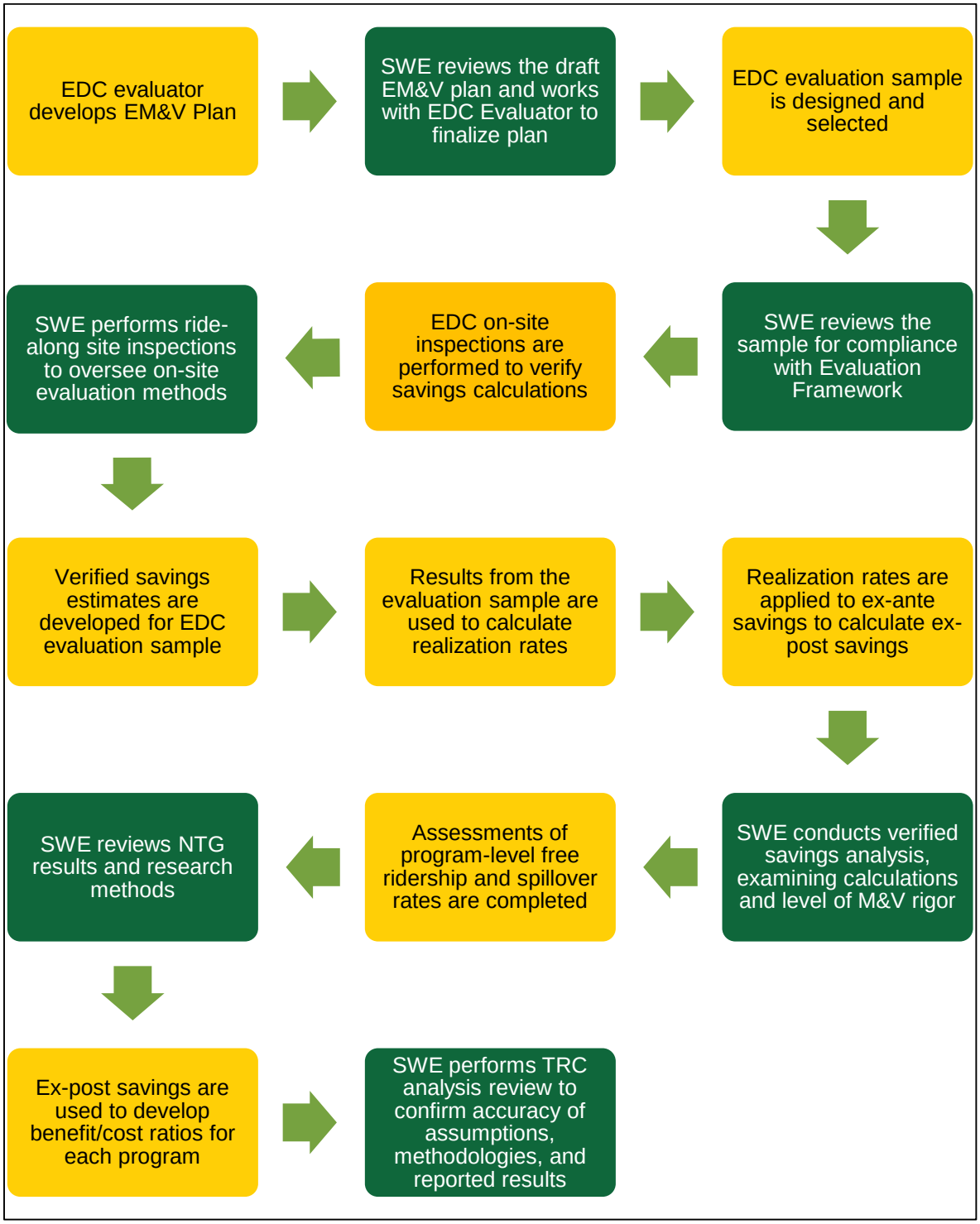
- **Amount of energy and demand savings.** More frequent gross impact evaluations are warranted for programs or initiatives that are expected to produce the most energy and demand savings; conversely, programs and initiatives with low savings levels may not warrant annual gross savings evaluations. In Phase III, behavioral programs are expected to generate a lot of savings and hence warrant annual impact evaluations. Energy efficiency education programs in schools, in contrast, account for lower savings levels and do not merit annual gross impact evaluations.
- **Program continuity / discontinuity.** Programs and initiatives undergoing changes in 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. Appliance retirement programs are an example of programs undergoing a change in implementation contractors and are therefore in need of gross impact, net impact, and process evaluations; if realization rates differ markedly from those measured in Phase II, then more frequent evaluations would be prudent. Small business direct-install programs are an example of programs that often have few changes in realization rates, incentive levels, or other factors, and therefore may not need annual evaluations.
- **Market or technology continuity / discontinuity.** Rapid change in a market warrants more frequent evaluations of a program or initiative targeting that market. A clear example is upstream lighting, given the disruption to the market brought on by EISA and the rapid development of LED technology; in this case, gross savings parameters may be relatively stable and hence not require annual updates, while net savings could be expected to change more rapidly and hence warrant more frequent measurement.
- **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.
- **Uncertainty and the risk of being wrong.** Based on all of the above, the greater the risk of over- or under-estimating savings, the greater the need for a gross impact evaluation.
- **Underperforming expectations.** If realization rates are lower than expected, then a process evaluation may be required to assess the causes of the shortfall.

Each EDC should use the above criteria to propose preliminary five-year evaluation schedules for every program and initiative; the proposed schedules will be reviewed by the SWE. The EDC EM&V plans should include the rationales for the schedule for each program and initiative. Note that reducing the frequency of some evaluations does not necessarily equate with conducting fewer evaluations overall. In particular, sampling requirements (as discussed in Section 3.6) may necessitate larger samples than some EDCs have relied on in the past.

Section 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 evaluation contractor. 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. Figure 6 shows the specific SWE audit activities and their correspondence to the evaluation steps.

Figure 6: 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 evaluation contractors to implement corrective actions within the program year.

4.1 EDC Report and SWE Report Schedule

The semi-annual and annual reports defined by the PUC are one of the ways by which stakeholders are informed about the spending and savings impacts of Act 129 EE&C plans. These semi-annual and 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 III.

4.1.1 EDC Report Schedule

The EDCs are required to submit semi-annual and annual reports to the SWE Team and the TUS. In the *Phase III Implementation Order* entered June 11, 2015, the PUC noted that Act 129 requires EDCs to submit an annual report documenting the effectiveness of their EE&C plans, measurement and verification of energy savings, evaluation of the cost-effectiveness of their expenditures, and any other information the PUC requires.[93]

The SWE Team provides the EDCs with semi-annual and annual report templates, which are available on the PA Act 129 SharePoint Site. The deadlines for the EDC reports are provided in Table 19.

Table 19: EDC Reporting Schedule

Report	Due	Savings Reported
Program Year X, Semi-Annual Report #1	January 15	• EE participation and impacts from Q1-Q2 • Gross verified demand response impacts (PYX and Phase III) • Implementation and evaluation updates • Gross reported savings PYTD • Sum of Incremental Annual Phase III savings (progress towards goals)
Program Year X, Semi-Annual Report #2 - Preliminary Annual Report	July 15	• EE participation and impacts from Q3-Q4 • Implementation and evaluation updates • Gross reported savings PYTD • Sum of Incremental Annual Phase III savings (progress towards goals)

[93] Implementation Order issued June 11, 2015, at Docket No. M-2014-2424864

Report	Due	Savings Reported
Program Year X – Final Annual Report	November 15	• Impact evaluation results (realization rates and confidence intervals) • Gross verified EE savings (PYX) • Gross verified demand response 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 III 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 estimates filed in their EE&C plans. Requested items will include:

- 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 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 annual reports if the change in annual portfolio savings reported by an EDC is less than 1%. In instances where the change is greater than 1%, 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 by the EDC.

4.1.2 Statewide Evaluator Report Schedule

In Phase III, the SWE Team will submit two reports to the PUC each program year. By August 15, an Update Report will be prepared and will include the gross verified demand response impacts for the program year and a summary of reported EE savings for the program year along with the results of the SWE ex ante audit. By February 28 (nine months after the close of the program year) an Annual Report will be submitted and 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 measurement and verification 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 PEG meetings and/or a summary of recently issued guidance memos.

The deadlines for the SWE reports to the PUC are presented in Table 20.

Table 20: SWE Reporting Schedule

Report	Due	Savings Reported
DRAFT Program Year X, Update Report	August 1	• Summary of EDC-verified DR impacts • SWE DR 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, Update Report	August 15	• Final update report; comments from TUS staff and EDCs addressed
DRAFT Program Year X Report	January 15	• Summary of EDC gross verified savings claims from EE and DR 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 Report	February 28	• 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 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.

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 data request response for Q1-Q2 should equal the PYTD kWh savings for that program in the EDC semi-annual report to the PUC. Additionally, the data request includes instructions for uploading the data requested to the EDC-specific Act 129 SharePoint site page. 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.
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.[94] 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 or D).

4.2.1.1 Desk Audits

As part of its contract with the Pennsylvania PUC, the SWE will complete desk audits for the nonresidential, low-income, residential lighting, residential appliance rebate, residential appliance recycling, and residential new construction 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

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

independently calculate savings and verify them against the savings reported by the EDC.

2. **Semi-annual and annual report reviews** through which the SWE will verify that the values presented in EDC semi-annual and 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 provide confidence in the gross verified program savings documented in the EDC annual reports, and transparency in the evaluation process. The SWE will present the findings and recommendations from its annual audit activities in its annual report for each program year. 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 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 evaluation contractor will present the savings impacts using both the TRM savings protocol and the protocol that the EDC's evaluation contractor believes is more appropriate for the measure. The SWE will review the savings protocol proposed by the EDC's evaluator and provide a recommendation to the PUC to 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.

The majority 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 evaluation contractors. 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 prospective recommendations will be to enhance program delivery and cost-effectiveness and improve the accuracy of savings protocols used by the ICSPs and EDC evaluation contractors.

4.3.1 Survey Instrument Review

Participant surveys are the most common form of data gathering used by EDC evaluation contractors 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 III, the evaluation contractors must submit draft survey instruments (in advance of survey implementation) that include process 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 considered to be 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 evaluation contractors once survey instruments have been uploaded to the SWE Team SharePoint site for review. The SWE will provide comments and suggest any possible revisions within five business days. Evaluators are not required to change the survey instruments based on the SWE's feedback, but 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 will 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 15 days after the submittal of the EDC's final annual report for a program year. This request includes only the ex post savings analysis the EDC evaluation contractor used to calculate gross 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." The three components of the SWE annual data request are presented below.

4.3.2.1 Evaluation Sample Population

For each program or initiative, the evaluation contractor should provide a table that contains the following information for each project in the completed evaluation sample. 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, in which stratum was this project located?
- Selection Type: When the sample was designed, was this project a primary sample or an alternate?
- Evaluation Activity: What type of evaluation activity was performed in order to develop verified savings estimates for this project (e.g., phone interview, online survey, desk review, site inspection, building simulation, or multiple methods)?
- M&V Approach: Which approach was used to calculate the verified savings for this project (e.g., simple verification, IPMVP Option A-D, or other appropriate methodology)?
- Meters Deployed: Was any type of logging equipment deployed at this site to collect information on key parameters in the savings calculations? (Yes/No)
- Verified kWh/yr: What are the verified annual kWh/yr savings for the project?
- Verified kW: What are the verified peak kW savings for the project?

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. For programs that utilize a regression-based analysis of monthly utility bills for an attempted census of participants, evaluators should provide the analysis data set used to estimate savings along with 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 as well as the relevant regression output, such as:

- Number of observations used, number of missing values
- ANOVA table with degrees of freedom, F-value, and p-value
- R-square and adjusted R-square 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 (VIFs) or other tests for multicollinearity

4.3.2.2 Evaluation Sample Audit

The SWE will select a sample of projects from each evaluation group provided in response to Section 4.3.2.1 and provide the EDC evaluation contractor with a list of the Unique Identifiers (UI) for those projects. Within 15 days of receiving the UIs, EDC evaluators must provide the evaluation documentation and findings for each project. The SWE will conduct a desk audit of these projects to confirm the reliability of the savings estimates. There is additional detail regarding these SWE desk audits in Section 4.3.4.

The documentation and findings to be supplied by the EDC evaluation contractor will vary per the evaluation approach they used. These items should include:

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

4.3.2.3 TRC Model Audit

The evaluation contractor 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 benefit/cost ratio. Key inputs the SWE will examine include:

- Discount rate
- Line loss factors
- Avoided costs of generation energy and capacity as well as T&D avoided costs
- Incremental measure costs
- 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

The SWE will present the findings and recommendations from its annual audit activities in its 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 EDC's 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 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 evaluation contractor will present the savings impacts using both the TRM savings protocol and the protocol deemed more appropriate for the measure. The SWE will review the savings protocol proposed by the EDC evaluator and provide a recommendation to the PUC to approve or reject the protocol. The SWE's recommendation should not be construed as PUC approval, as the PUC has the ultimate authority to approve or reject the savings calculated using the proposed protocol.

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 prospective recommendations will be to enhance program delivery and cost-effectiveness and improve the accuracy of savings protocols used by the ICSPs and EDC evaluation contractors.

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 evaluation contractors' sampling approaches at three stages during program evaluation.

1. **Evaluation, Measurement, and Verification (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. The EDCs are encouraged to submit the sample design before the EM&V plan to expedite the SWE's approval.
2. **Quarter 3 of the Program Year** – Within a month of the close of Q3 (i.e., by May 15) for each program year, evaluation contractors should submit an updated sampling plan for each EDC program. 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 evaluation contractor'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 annual report will contain a recommendation to focus more heavily on that technology the following year. 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. 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 Desk Audits

The SWE audit of the EDC evaluations will include all review activities required to assess the quality control, accuracy, and uncertainty of verified savings estimates. Annually, the

SWE Team will request verification data for projects in the sample drawn by the EDC evaluation contractor for each EDC program as described in Section 4.3.2.2. Typically, projects for the SWE Evaluation Sample Audit will be selected after the EDC annual report has been filed from the evaluation sample population submitted as part of the SWE Annual Data Request. If an evaluation contractor completes a significant share of the verified savings analyses for a program year in advance of the reporting deadline (November 30), the SWE will consider a two-stage sampling process to allow increased discussion prior to the inclusion of audit findings in the SWE Annual Report. The SWE will audit the M&V methods used by the evaluator to ensure the verified savings are calculated using approved protocols.

The SWE will review the evaluation processes and compare them with the approved evaluation plans. In addition, for quality assurance, the audit activities will include some ex ante savings checks such as: a review of randomly selected incentive applications, verification of the proper application of TRM assumptions, and assessment of the consistency of data between incentive applications and the EDC data tracking system. The evaluation reports requested from the EDC evaluation contractor should include the following information:

- Site-specific M&V plans (applicable only to commercial and industrial programs), clearly showing the data collection process and how it is utilized in savings analysis
- Site inspection findings (applicable to all programs)
- Description of metering methods, including measurement equipment type, location of metering equipment, equipment set-up process, photographs of meter installation, metering duration for which data were collected, metered results, and accuracy of the results
- Savings calculations, with all supporting information
- Incentive applications
- Other pertinent information

In its annual reports, the SWE will document findings and recommendations resulting from these desk audits, as well as actions taken by EDCs to address them. If an EDC evaluation contractor submits verified savings analyses for audit before the November 30 due date, the SWE will work to provide audit findings and recommendations to the EDCs for review and discussion prior to documenting them in the SWE's annual report.

4.3.5 Site Inspections

Site inspections are essential for the accurate evaluation of programs and will represent a significant portion of the EDCs' evaluation efforts for residential and nonresidential programs.[95] Because of the importance of this task, the SWE Team will work closely with the EDCs to ensure that site inspections are planned and executed carefully and that site

[95] SWE site inspections are typically focused on large nonresidential projects, but may include a small number of site visits for low-income or residential whole-home programs in Phase III.

inspectors have the appropriate experience and training. The SWE Team will audit the following steps in each EDC's site inspection process:

- Training of site inspectors to collect site-specific information
- Random sampling of projects
- Development of the evaluation tracking database and site inspection forms
- Grouping of inspections by geographic location (as appropriate) to minimize time allocation, labor, and direct costs associated with conducting inspections
- Contacting sites prior to any visit to ensure availability and to ensure the resident or facility staff is not "surprised" by the visit
- Performing site inspections and entering all required data into the program evaluation database.

In general, the SWE audit activities will fall into two categories:

1. **Ride-Along Site Inspections (Audits):** The SWE may perform "ride-along audits," in which the SWE accompanies the EDC evaluator on a site inspection to validate and confirm that EDC evaluators are using approved protocols when performing evaluation activities. This includes checking for adherence with the TRM, where applicable, and compliance with the SWE Evaluation Framework. 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.
2. **Independent Site Inspections (Audits):** Although less frequent than ride-along audits, the SWE may 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. The SWE will conduct relatively fewer independent site inspections than ride-along inspections. The SWE expects to conduct more independent inspections at the beginning of each Phase and then fewer such inspections as it becomes more confident that the ICSPs' reported savings estimates and evaluation contractors' verification activities are accurate. 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 is committed to working collaboratively with the EDCs and the EDC evaluators to conduct audit activities and ensure the accuracy of ex ante savings and realization rates that support unbiased estimations of verified gross energy and demand impacts for the Act 129 programs.

The SWE will produce and distribute its ride-along site inspection reports (RA-SIRs) and independent site inspection reports (I-SIRs) to EDC evaluators within 15 business days of completing a ride-along to document its site inspection findings and verified savings calculations. In the case of ride-along inspections, the EDC evaluation contractors will calculate verified savings and SWE inspectors will verify them. Findings and recommendations resulting from RA-SIRs and I-SIRs, as well as actions taken by EDCs to address the findings and recommendations, will be documented in the SWE annual reports.

1. **Ride-Along Site Inspection Reports:** RA-SIRs will focus on process findings that also may affect the gross impacts verified by the evaluation contractors. The SWE also will review evaluators' site inspection reports to ensure that all savings calculations and critical site findings have been identified. The RA-SIRs will be completed after the EDC evaluators have shared their site inspection reports and engineering calculations with the SWE. EDC evaluators will have the opportunity to review RA-SIRs and discuss key issues and/or discrepancies with the SWE. Resolutions will be reached collaboratively by the SWE and the EDC evaluators.
2. **Independent Site Inspection Reports:** If an independent site inspection is completed by the SWE, I-SIRs will include process findings related to program delivery and an independent SWE assessment of ex ante project impacts. The SWE will calculate verified savings for all independent inspection samples. Because independent site inspections are conducted on sites not selected by the EDC evaluation contractors, I-SIRs will be issued shortly after SWE evaluation activities have been completed.

If the SWE Team elects to conduct an independent site inspection, the EDC and evaluation contractor will be notified well in advance of the visit. Verified savings estimates from projects receiving a SWE I-SIR can be included in the gross impact evaluation sample and subsequent realization rate calculation at the discretion of the EDC evaluation contractor. EDC evaluators will not be required to incorporate the results from I-SIRs in the final realization rate calculations. As appropriate and with substantial justification, the SWE will request further quarterly and annual information on specific observations made during independent site inspections. The EDC evaluators will be responsible to address the SWE's independent observations in a timely manner.

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 appliance retirement programs (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 in the SWE Net-to-Gross Study Methods guidance document distributed to the EDCs and their evaluation contractors on February 27, 2012, which remains the document in effect for programs not addressed by the Evaluation Framework. The levels of rigor (basic, standard, and enhanced) and the methods involved in each are outlined in Table 21.

Table 21: Rigor Levels Adapted from the California Energy Efficiency 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 nonparticipants • 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 sections 3.6 and 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 evaluation contractors 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 the 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 evaluation contractors 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 that the:

- Research objectives are complete relative to the type of process or market evaluation planned.
- Sample design is sufficient and appropriate to address the objectives.
- Data collection approaches are appropriate and executed per plan.
- Data collection instruments address the objectives and do not introduce bias.
- Analysis and report writing convey the findings clearly and draw reasonable conclusions.
- Recommendations are actionable and clearly identify which parties should address the recommendation.
- EDCs follow up on process evaluation recommendations and report to the SWE the action 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 skill 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 management's 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' interaction 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 too low a level of market effects, free drivers, or spillover
- Intended or unanticipated market effects

4.5.2 Sample design

Sampling for process and market evaluations should follow sampling approaches similar to 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 22 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 22: 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 it 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 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 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, nonparticipants, 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 SWE standard set of demographic /firmographic questions (e.g., comparable to Census or industry data)
- If needed, request for permission to call back or email with follow-up questions (especially useful when conducting in-depth interviews); collection of appropriate call back information, best phone, email address, etc.
- Request for any additional comments from 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
- Are 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 is asked of 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 (AAPOR).[96] 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 evaluation contractor 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. These findings will be reported in the SWE Annual Report.

- The SWE review of process findings for these various programs by EDC will help to identify best practices across the state.
- The SWE also will compare process evaluation findings to process and delivery strategies of similar best programs throughout the United States.

[96] See http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/Standard_Definitions_07_08_Final.pdf.

- The SWE will present the findings in a manner that highlights areas of success within the portfolio of EDC projects and that identifies areas of improvement.
- The SWE also will report on selected EDC responses to the recommendations.

4.6 Cost-Effectiveness Evaluation Audit

The SWE cost-effectiveness assessment will include a review of the benefit/cost (B/C) ratio formulas, benefits, costs, and TRC ratios at the EDC project level, EDC program level, and EDC plan level. The SWE will determine whether TRC calculations have been performed according to the PUC's latest TRC 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 B/C ratio, calculation formulas, and TRC outputs.

4.6.2 Inputs and Assumptions

Key inputs and assumptions the SWE will examine include:

- Discount rate
- Line loss factors
- Avoided costs of energy and capacity
- Incremental measure costs
- 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:

- A review for TRC Order compliance regarding:
    - Formulas
    - Benefits
    - Costs
    - Utility avoided costs assumptions

- A review of EDC accounting practices, including:
    - Division of costs and benefits between programs
    - Appreciation/depreciation rates

For Phase III, EDCs may choose to adopt a proprietary benefit-cost software product for their TRC analysis. For EDCs using proprietary products, the SWE 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.

Section 5 Resources and Meetings

This Evaluation Framework is intended to serve as a resource for EDC program administrators and evaluation contractors. The Framework is a living document and will be updated annually in Phase III, 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 will provide documents for sharing on the PUC's public website,[97] which provides information to interested stakeholders on the actual kWh/yr 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 created a PA Act 129 SharePoint site to improve communication and coordination of activities among the SWE Team, the TUS, the EDCs and their evaluator contractors, and the Energy Association. 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 seven 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-2015-3).

The PA Act 129 SharePoint includes:

- **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 the following features: calendar, task lists, technical libraries, report libraries, submission logs, and discussion boards.
- **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, track progress, and store lists of unresolved issues.
- **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 III, the SWE will create Level 1 folders in each individual EDC site and the common SWE site for each program year, and Level 2 folders to house documents such as

[97] The URL for the Act 129 directory of the PUC's website: http://www.puc.pa.gov/filing_resources/issues_laws_regulations/act_129_information.aspx

reports, tracking data, and data requests/responses. The Level 1 and 2 folder structure will be consistent across the individual EDC sites. The common SWE site will house PEG meeting minutes and agendas, a data request tracking sheet(s), as well as the final versions of the SWE reports. Additionally, the common SWE site will maintain all of the SWE guidance memos, the master contact list, approved IMPs, guidance memos, study memos, and a calendar with important dates.

5.3 Program Evaluation Group Meetings

The SWE will chair and set the agenda for quarterly meetings of the PEG and will prepare minutes of these meetings. These meetings will be conducted per the same format used during Phase II of Act 129.

5.4 Stakeholder Meetings

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

Section 6 Measure-Specific Evaluation Protocols (MEPs)

This section provides additional guidance and measure-specific evaluation protocols to estimate energy and demand savings associated with behavioral modification and demand response programs.

6.1 Behavioral Conservation Programs

Behavioral conservation programs such as Home Energy Report (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 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 two most prevalent behavior-based programs in the Phase III EE&C plans approved by the PUC; Home Energy Reports and Business Energy Reports. If EDCs chose 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.

This protocol does not address behavioral demand response. Guidelines for evaluation of behavioral demand response programs are addressed in the Demand Response protocols of the Evaluation Framework.

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 included along with energy savings when calculating benefits for the TRC test.

6.1.1.1 Experimental Design

Act 129 HER and BER programs must 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 where an RED approach is needed because only a subset of the homes encouraged to visit the web portal will actually do so.

The SWE's review of Phase III EE&C 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 net-to-gross 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 Business Energy Report 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 evaluation contractors 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.

6.1.1.1.1 Group Sizes

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

1. 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 III EE&C budgets and projected savings, are measured in 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 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 in this example was 150 kWh per household annually, the relative precision at the 85% confidence level would be:

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

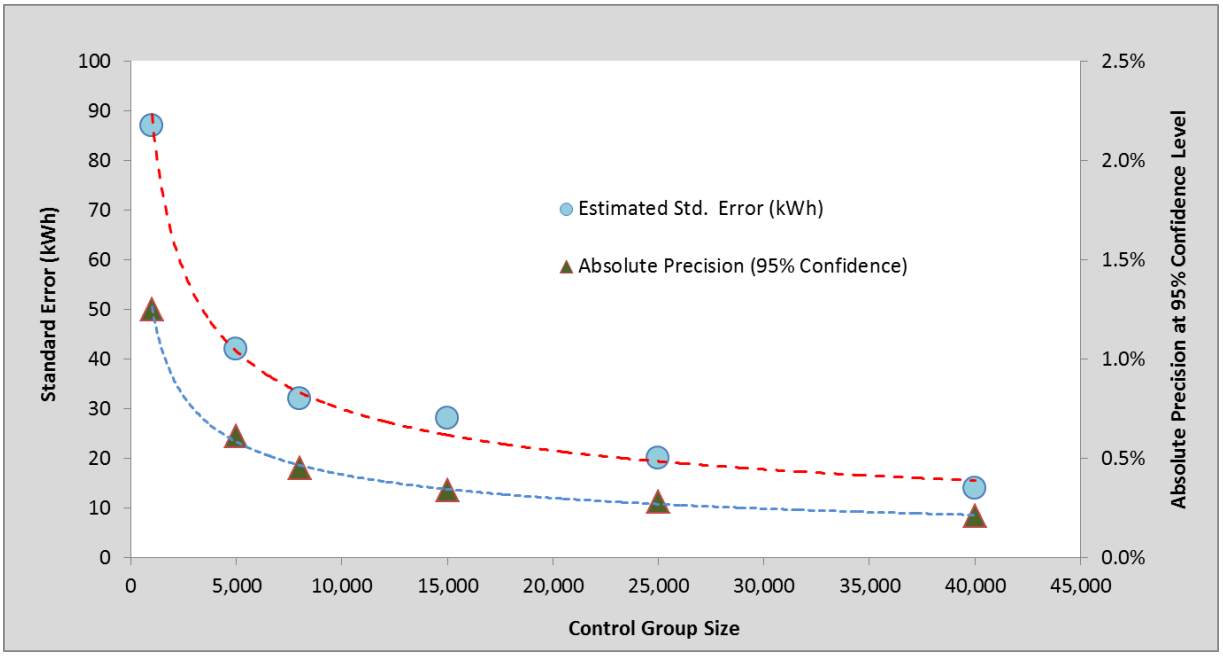
Extremely large control group sizes would be necessary to achieve ±15% relative precision at the 85% confidence level. For BER programs where customer size and consumption patterns are highly variable and expected percent impacts are smaller, 85/15 is likely impossible.

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 evaluation contractors 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 evaluation contractors 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 evaluation contractors 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 7 shows the expected output from group size investigation (either method). As the control group sizes increase, the expected standard error shrinks and the expected precision improves.

Figure 7: Hypothetical Sample Size Simulation Output



The relationship is non-linear, and this creates a "diminishing returns" effect 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 evaluation contractors should never draw samples of homes from the treatment and control groups for gross energy efficiency 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 event to some extent for demand reduction analysis, is acceptable.

6.1.1.1.2 Opt-Outs and Account Closures

Over time, some homes and businesses assigned to behavioral conservation programs will close their account 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 evaluation contractor.

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 evaluation contractors 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 Uniform Methods Project Residential Behavior Protocol[98] 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.

6.1.1.1.3 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, evaluation contractors must still verify that the randomization process was successful, as described in Section 6.1.1.3.

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 or to expand the program to additional participants. 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 to calculate MWh impacts, the aggregate MWh savings across cohorts can be summed to calculate program performance. EDC evaluation contractors can perform a weighted average calculation to produce relevant statistics, such as the average annual kWh savings per home or average percent savings per home, using the number of active treatment group homes as the weighting factor.

6.1.1.3 Equivalence Testing

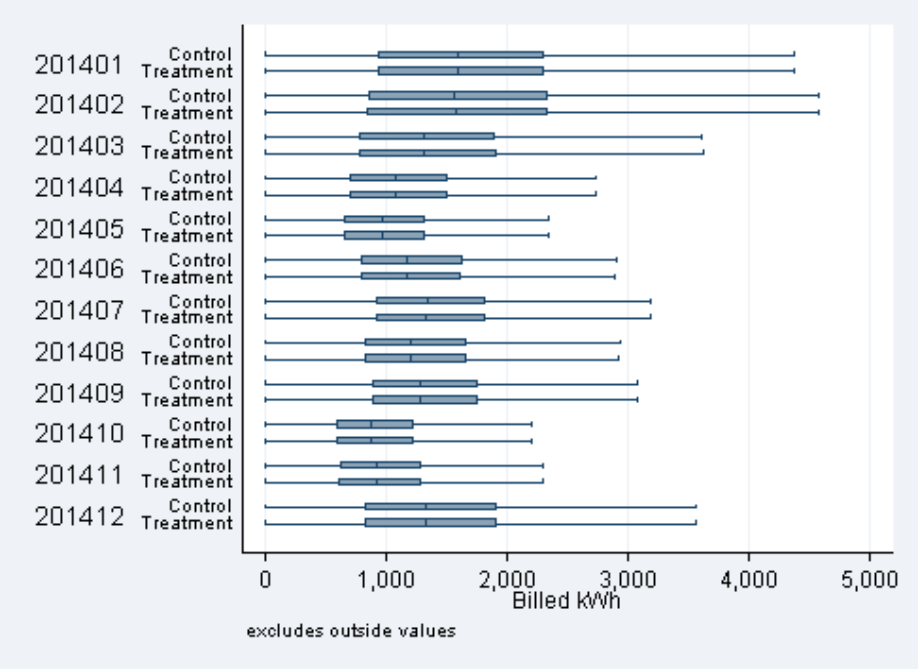
Validation of the pre-treatment equivalence of the treatment and control groups is an important feature of behavioral program evaluation because randomization is so critical to

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

the ability to develop unbiased measurements of behavioral program impacts. Randomization can be performed by the EDC, the behavioral program vendor, the EDC evaluation contractor, or the SWE (if requested). Regardless of who performs the randomization, EDC evaluation contractors 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 8 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 8: 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 the randomization sound, EDC evaluation contractors 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. 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 evaluation contractors should alert the SWE immediately to discuss the appropriate corrective action.

When the randomization is compromised and the treatment has begun, the SWE will work with the EDC evaluation contractor 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).

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, the resulting estimate of the treatment effect will 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 AMI/AMR in place for all customers and the capability to provide that data to evaluation contractors 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 evaluation contractor 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 are a function of meter read dates and vary across accounts. Since the interval between meter readings varies by customer and by month, EDC evaluation contractors 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 actually 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 evaluation contractor for analysis. When such data is calendarized for analysis, evaluation contractors 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 23 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 23: Estimated Meter Read Calendarization Example

Meter Read Date	Days in Cycle	Estimated or Actual	Billed kWh	Average Daily kWh
2/15/2015	30	Actual	1,500	50
3/15/2015	28	Estimated	1,100	38.2
4/15/2015	31	Estimated	900	
5/15/2015	30	Actual	1,400	

6.1.1.4.1 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" regression, LFER models estimate the average treatment effect 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.[99]

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

[99] https://www4.eere.energy.gov/seeaction/system/files/documents/emv_behaviorbased_eeprograms.pdf

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 average treatment effect on an absolute basis (kWh).

3) **Lagged Seasonal (LS) Model**. This model is similar to the LDV, but uses pre-treatment consumption data for each home slightly differently. Instead of creating a single lag term, the lagged seasonal model contains three lag variables: one for average usage (all months), one for average summer usage, and one for average winter usage. The LS model estimates the average treatment effect on an absolute basis (kWh).

4) **Natural Log Panel Regression (Log) Model**. This model is similar to the LFER model except that it uses the natural log of consumption as the dependent variable rather than a level consumption term. By using the natural log of kWh, the Log model produces a "difference-in-differences" calculation on a relative (%) rather than an absolute basis. This approach normalizes customer size and provides the average percent savings per participant.

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. In order to avoid the temptation of estimating multiple models and selecting the approach with the most favorable savings estimate, EDC evaluation contractors must specify the model specification that will be utilized to calculate savings in their EM&V plans and provide justifications for their choice.

The LFER, LDV, and LS models are all acceptable for estimating and claiming verified energy savings from Act 129 behavioral programs. Evaluation contractors are encouraged to consider the Log model as a supplemental analysis, but it should not be used to estimate and claim savings. While this is an interesting research question and worthwhile for EDCs to investigate, the Log model does not answer the fundamental EM&V question of "What were the kWh savings achieved by the program?" Because the Log model normalizes customer size, the ability to detect and account for different savings levels by customer size is lost. Applying the average percent reduction to the average customer size will not produce accurate energy savings estimates if small and large consumers save energy at different rates. If the Log model is going to be considered, evaluation contractors should note this in the EM&V plan in addition to the specification of the primary model that will be used to claim energy savings.

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, evaluation contractors 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.

6.1.1.5.1 Technical Guidance on Behavioral Models

The basic form of the LFER model is shown in Equation 13. 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 household-level fixed effects:

Equation 13: Fixed Effects Model Specification

$$\mathrm{kWh}_{\mathrm{imy}} = \beta_{\mathrm{i}} + \sum_{\mathrm{m}=1}^{12}\sum_{\mathrm{y}=2011}^{2021} \mathrm{I}_{\mathrm{my}} * \beta_{\mathrm{my}} + \tau_{my} * \sum_{\mathrm{m}=1}^{12}\sum_{\mathrm{y}=2011}^{2021} \mathrm{I}_{\mathrm{my}} * \mathrm{treatment}_{\mathrm{imy}} + \varepsilon_{\mathrm{imy}}$$

Table 24 defines the model terms and coefficients in Equation 13.

Table 24: LFER Model Definition of Terms

Variable	Definition
$\mathbf{kWh_{imy}}$	Customer i's average daily electric usage in month m of year y.
$\boldsymbol{\beta_i}$	The intercept term for customer i, or the "fixed effect" term. Equal to the mean daily energy use for each customer.
$\mathbf{I_{my}}$	An indicator variable that equals one during month m, year y, and zero otherwise. This variable models each month's deviation from average energy.
$\boldsymbol{\beta_{my}}$	The coefficient on the month-year indicator variable.
$\mathbf{treatment_{imy}}$	The treatment variable. Equal to one when the treatment is in effect for the treatment group. Zero otherwise. Always zero for the control group.
$\boldsymbol{\tau_{my}}$	The estimated treatment effect in kWh per day; the main parameter of interest. Estimated separately for each month and year
$\boldsymbol{\varepsilon_{imy}}$	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 group 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 14 shows the basic form of the LDV model. Unlike the LFER model specification, all accounts share a common intercept (β_0) 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 14: LDV Model Specification

$$kWh_{imy} = \beta_0 + \sum_{m=1}^{12}\sum_{y=2011}^{2021} \mathrm{I}_{my} * \beta_{my} + kWh_{i,m,y-n} * \beta_{m,y-n} + \sum_{m=1}^{12}\sum_{y=2011}^{2021} \mathrm{I}_{my} * \tau_{my} * \mathrm{treatment}_{imy} + \varepsilon_{imy}$$

Table 25 defines the model terms and coefficients in Equation 14.

Table 25: 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. The term n represents the number of years home i has been in the program. This term controls for variability in customer characteristics such as home size and heating fuel.
$\beta_{m,y-n}$	The coefficient on the home-specific pre-assignment usage term.
$\mathrm{treatment}_{imy}$	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.
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 evaluation contractors should use post-only models only when the treatment and control groups are balanced on usage and selection criteria.

Equation 15: Lagged Seasonal Model Specification

$$kWh_{imy} = \beta_0 + \sum_{m=1}^{12}\sum_{y=2011}^{2021} \mathrm{I}_{my} * \beta_{mys} * (AvgPre_i + AvePreSummer_i + AvePreWinter_i) + \sum_{m=1}^{12}\sum_{y=2011}^{2021} \mathrm{I}_{my} * \tau_{my} * \mathrm{treatment}_{imy} + \varepsilon_{imy}$$

Table 26 defines the model terms and coefficients in Equation 15.

Table 26: Lagged Seasonal 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.
$AvgPre_i$	Average daily usage for customer i in the pre-treatment period.
$AvePreSummer_i$	Average daily usage for customer i in the pre-treatment period during June through September.
$AvePreWinter_i$	Average daily usage for customer i in the pre-treatment period during December through March.
$\mathrm{treatment}_{imy}$	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.
ε_{imy}	The error term.

The lagged seasonal 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. This added precision can justify a slightly smaller control group size. In repeated simulations, the LS model will occasionally produce a result that significantly misrepresents the actual treatment effect. Like the LDV model, the LS model is poorly equipped for pre-treatment differences between the treatment and control groups. EDC evaluation contractors should use post-only models only when equivalence tests indicate that the randomization for a cohort is uncompromised.

Equation 16 provides the model specification for the natural log panel regression model.

Equation 16: Natural Log Panel Regression Model

$$\ln\_\mathrm{kWh}_{imy} = \beta_i + \sum_{m=1}^{12}\sum_{y=2011}^{2021} \mathrm{I}_{my} * \beta_{my} + \tau_{my} * \sum_{m=1}^{12}\sum_{y=2011}^{2021} \mathrm{I}_{my} * \mathrm{treatment}_{imy} + \varepsilon_{imy}$$

Table 27 defines the model terms and coefficients in Equation 16.

Table 27: Log Model Definition of Terms

Variable	Definition
$\mathbf{ln\_kWh_{imy}}$	The natural log of customer i's average daily electric usage in month m of year y.
$\boldsymbol{\beta_i}$	The intercept term for customer i, or the "fixed effect" term. Equal to the mean daily energy use for each customer.
$\mathbf{I_{my}}$	An indicator variable that equals one during month m, year y, and zero otherwise. This variable models each month's deviation from average energy.
$\boldsymbol{\beta_{my}}$	The coefficient on the month-year indicator variable.
$\mathbf{treatment_{imy}}$	The treatment variable. Equal to one when the treatment is in effect for the treatment group. Zero otherwise. Always zero for the control group.
$\boldsymbol{\tau_{my}}$	The estimated treatment effect in % reduction in electric consumption. Estimated separately for each month and year.
$\boldsymbol{\varepsilon_{imy}}$	The error term.

Like the LFER model, the Log model includes participant-level fixed effects that eliminate any time-invariant characteristics from the estimation. The Log model is less susceptible to structural differences in consumption patterns between treatment and control than the post-only models. Depending on the functional form of the treatment effect, the Log model can actually produce the most precise estimates of the four models discussed. However, it does not estimate program energy savings—it estimates only the percent savings of the average participant. While the difference is subtle, the merits of the Log model do not outweigh the fact that it fails to answer the primary research question for Act 129 evaluations. Therefore, if EDC evaluation contractors use the Log model, they will also have to run one of the other three models to provide the necessary estimates of energy savings.

Table 28 provides a summary of the strengths and weaknesses of the four classes of regression models discussed in this section.

Table 28: Summary of Model Pros and Cons

Model Specification	Advantages	Disadvantages
Linear Fixed-Effects Regression (LFER)	Best equipped to net out pre-treatment differences in energy consumption	Less precise because between-participant variation is not used
Lagged Dependent Variable (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
Lagged Seasonal Interaction (LS)	Most precise, on average	Occasionally produces erratic estimates
Natural Log Panel Regression (Log)	Can produce more precise estimates if the treatment effect is not normally distributed. Directly estimates percent savings.	Does not estimate program energy savings

6.1.1.5.2 Monthly Impact Estimates

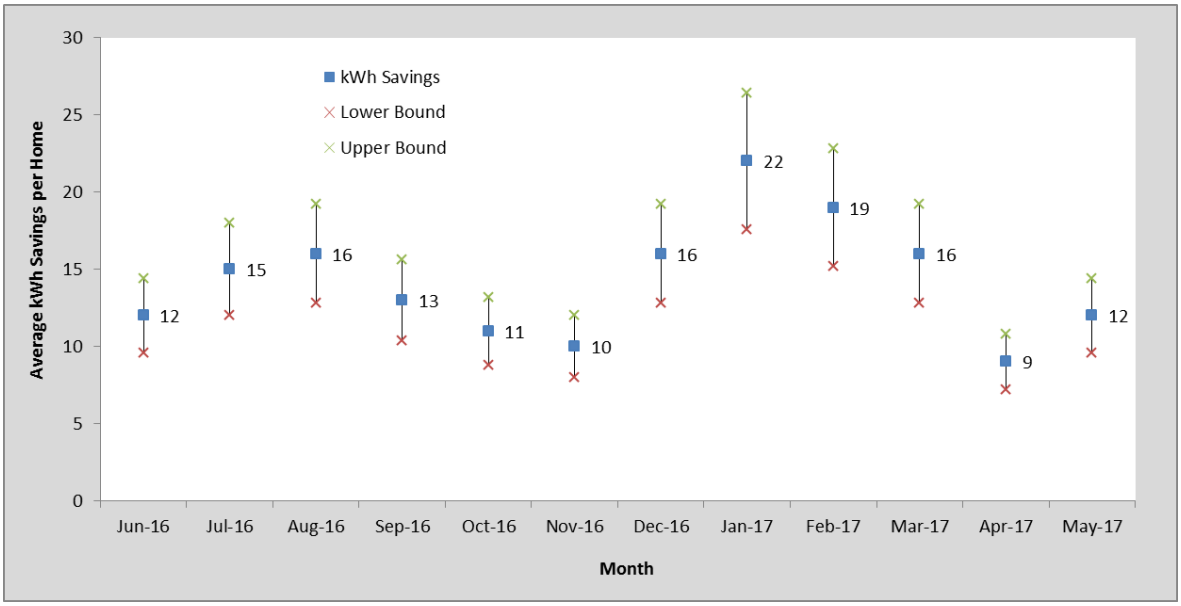
In each of the model specifications provided in Section 6.1.1.5.1, the parameter of interest (treatment) is interacted with an indicator variable (month dummies) to produce monthly estimates of the treatment effect (daily kWh savings). EDC evaluation contractors should use treatment/time dummy interaction 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 average treatment effect 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 Home Energy Report cohort began receiving HERs at the beginning of PY5 (June 2013). If, at the end of PY8 (May 2017), an EDC evaluation contractor estimated a cumulative regression model using a standalone treatment indicator variable, the coefficient would represent the average treatment effect for PY5, PY6, PY7, and PY8. If the treatment effect grew over time, which many evaluation studies have found, the PY8 savings from the program would be understated.

If evaluation contractors 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 Annual Reports should use graphics or tables like Figure 9 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 9: Monthly Impact Estimate Figure



EDCs should also consider presenting behavioral savings on a percentage basis. Percent impacts can be calculated using Equation 17 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.[100]

Equation 17: Percent Savings Calculation

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

6.1.1.5.3 Inclusion of Weather

The model specifications presented in Section 6.1.1.5.1 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. Although not necessary, weather variables can have significant explanatory power for electric consumption, and including them in the model may improve precision. EDC evaluation contractors 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.

[100] Alternatively, evaluation contractors could estimate the Log model and directly estimate percent reductions.

6.1.1.6 Peak Demand Impacts

The Pennsylvania TRM defines *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. 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. There are no peak demand reduction targets from energy efficiency in Phase III of Act 129 so there is a reduced emphasis on the accuracy of peak demand impacts compared to consumption reduction estimates. However, behavior-based offerings are expected to produce TRC Test results very close to 1.0 and precise peak demand impacts and time-differentiated energy savings do increase the accuracy of TRC Test results. 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.

6.1.1.6.1 With AMI

EDCs with hourly or sub-hourly meters on all of the program participants and the IT capabilities to retrieve the data for analysis have the ability to perform an actual ex-post analysis of demand impacts by comparing treatment and control group loads. The models described in Section 6.1.1.4.1 can, with a few adjustments, be used to estimate demand impacts. Average hourly demand (kW) becomes the dependent variable instead of average daily kWh. Monthly or program year dummy variables are used to estimate the average demand impact during the period of interest as opposed to the cumulative demand impact since the inception of treatment.

Data volume can become a constraint for EDC staff tasked with pulling interval data or evaluation contractors tasked with processing the data for analysis. Applying filters as early in the process as possible will help:

- Limit the data set to June, July, and August
- Exclude Saturdays, Sundays, and holidays
- Select records only from hours ending 15 through 18

If data management still proves burdensome to EDC staff and evaluation contractors, 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 evaluation contractors should notify the SWE to determine an acceptable degree of sampling based on the limitations in place.

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 8760 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 for use in the TRC Test.

6.1.1.6.2 Without AMI

EDCs without a fully deployed AMI system and the IT infrastructure to retrieve historic data for analysis will need to utilize an alternative approach to estimate peak demand impacts. Monthly billing data is far too coarse to measure peak demand impacts empirically. Instead, EDCs should take the measured annual energy savings (kWh) and allocate them across an 8760 load shape to estimate load reduction observed in each hour of the year. EDC evaluators should then average the impacts over the hours in the Act 129 peak demand definition. The selected load shape(s) should be mapped to the rate class of customers participating in the program and specific to the EDC service territory.

Evaluators should compare the distribution of monthly impact estimates provided by the regression analysis to the results of a premise-level 8760 load shape allocation. If it appears that savings are being understated in some months and overstated in others, it may be more accurate to select an end-use load shape or shapes that better align with observed monthly impacts and calculate peak demands and time-differentiated energy savings using those end-use load shapes.

6.1.1.7 Aggregate Impacts

Calculation of aggregate MWh or MW impacts from behavioral programs is conceptually straightforward and shown in Equation 18. Starting with the average treatment effect τ (measured in kWh/day and estimated separately by month), EDC evaluation contractors simply multiply by the number of days in each month and the number of active homes in the treatment group during the month.

Equation 18: Aggregate Impact Estimates

$$MWh\ Saved_{PY8} = \sum_{m=1}^{12} \tau_{my} * Days_{my} * Tx\ Accounts_{my}$$

Aggregate impacts should be calculated separately for each cohort in a behavioral program and then summed to arrive at an estimate of program performance. 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:[101]

> *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.

6.1.1.8.1 Downstream Programs

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

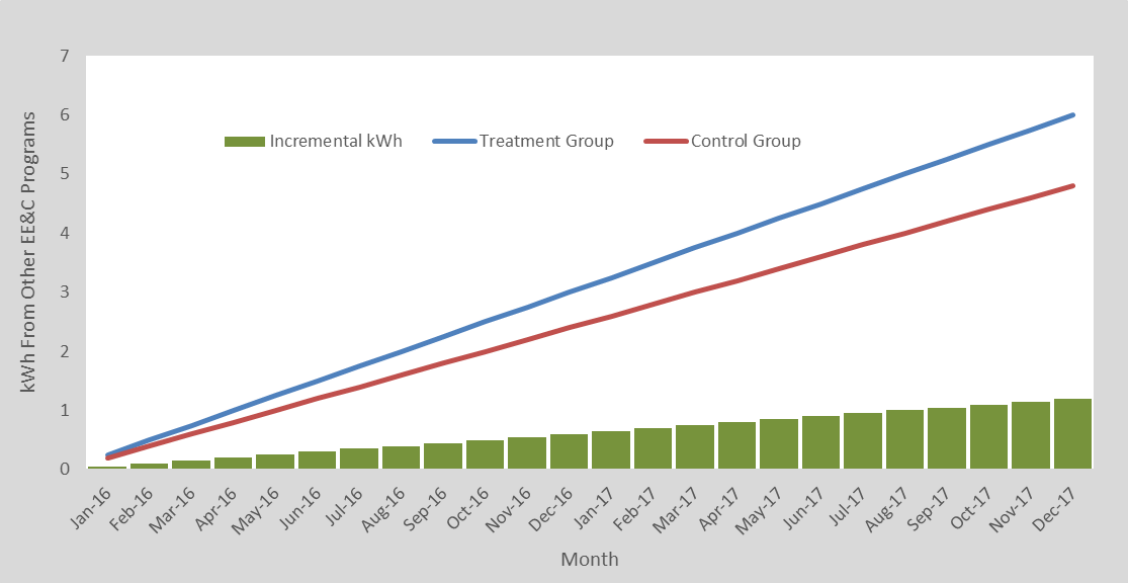
1) Match the program tracking data to the treatment and control homes by a unique identifier.
2) Assign each transaction to a month based on the participation date field in the tracking data.
3) Exclude any installations that occurred prior to the home being assigned to the treatment or control group.
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[102]. Evaluation contractors 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 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 10 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

[101] https://www4.eere.energy.gov/seeaction/system/files/documents/emv_behaviorbased_eeprograms.pdf (p. 31).
[102] 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.

separation in the average kWh savings per home. This difference, or incremental kWh, is what must be deducted from the behavioral programs' impacts to avoid double-counting.

Figure 10: Dual Participation Analysis Output



Dual participation analysis should be performed separately by cohort and magnitude of the adjustment reported in EDC Final Annual Reports. A long history of tracking data will be needed for cohorts that have been receiving treatment since Phase I or Phase II of Act 129. If an HER cohort began treatment in January 2012, EDC evaluation contractors would need program tracking data and evaluation results for all residential programs back to PY4 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 III of Act 129, long-running HER cohorts will begin to see dual participation savings from Phase I or Phase II that reach the end of their useful lives. Consider a measure with an EUL of five years installed in 2012. By 2018, the installed appliance has reached the end of its mechanical life and is no longer producing energy savings. EDC evaluation contractors 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.

6.1.1.8.2 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 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." EDC evaluation contractors are encouraged to perform surveys to estimate dual participation savings from upstream programs. If surveys are planned as part of the process evaluation, adding questions to explore this topic may be useful.

If EDC evaluators wish to allocate evaluation resources elsewhere, Table 29 provides default values that can be used to calculate a dual participation adjustment factor for upstream offerings. To account for the growing separation between the treatment and control groups over time, Table 29 relies on a conditional lookup based on the number of years since cohort inception to calculate the reduction factor. A "ceiling" is provided at year 4 to account for CFLs (which made up a large part of Phase I and Phase II upstream sales) reaching the end of their useful life.

Table 29: Default Upstream Adjustment Factors[103]

Years Since Cohort Inception	Default Upstream Reduction Factor
1	0.75%
2	1.5%
3	2.25%
4 and beyond	3.0%

The adjustment factors in Table 29 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 evaluation contractor's discretion. The following example shows a sample calculation for an HER program cohort in its third year.

$$PY8\ Average\ Impact\ per\ Home = 220\ kWh$$

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

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

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 evaluation contractors should apply the adjustment factors shown in Table 29 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

Phase III energy savings goals are based on incremental annual accounting of performance. Each program year, the new first-year savings achieved by an EE&C program

[103] 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. http://www.calmac.org/publications/2012_PGE_OPOWER_Home_Energy_Reports__4-25-2013_CALMAC_ID_PGE0329.01.pdf A 2014 Puget Sound evaluation found values lower than those in this table. https://conduitnw.org/_layouts/Conduit/FileHandler.ashx?RID=2963.

are added to an EDC's progress toward compliance. Unlike Phase I and Phase II of Act 129, whether or not a measure reaches the end of its EUL before the end of the phase does not impact compliance savings.

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.

The status quo in Pennsylvania's Act 129 programs has been to assume a one-year EUL. This perspective considerably simplifies accounting because the EDCs simply measure the savings at the meter each year the program operates and compare the benefits created to the costs of delivering the program. This perspective works favorably with incremental annual accounting because each year the program operates, new first-year (compliance) savings are gathered.

Industry studies, including one in Pennsylvania[104] during Phase II, have examined the appropriateness of a one-year EUL for behavioral programs and found evidence that the treatment effects persist for longer than one year after EDCs stop distributing them. Other works have found persistent savings three years after program cessation, indicating that a longer EUL may be appropriate.[105] To date, the PUC has not prescribed the measure life for behavioral programs and has identified persistence of behavioral savings as an area of investigation for the Phase III SWE team to inform targets and reporting protocols for future Phases of Act 129. Unless an alternative EUL was submitted and approved in a Phase III EE&C plan, EDCs should report annual savings consistent with the status-quo assumed one-year measure life.

[104] http://www.puc.state.pa.us/Electric/pdf/Act129/SWE_Res_Behavioral_Program-Persistence_Study.pdf
[105] http://www.energizect.com/sites/default/files/R32%20-%20Persistence%20of%20Eversource%20HER%20Pgm_Final%20Report%2C%203.30.16.pdf

Table 30 illustrates the interplay between EUL and compliance savings for two EDCs who run identical programs, but have different EUL assumptions approved in their Phase III EE&C plans. HERs are mailed to the same homes for all five years, and the average treatment effect is 150 kWh per home each year. Notice how EDC #2 claims no savings in PY9 or PY10 because the measured savings are not first-year incremental savings.[106] Rather, the measured savings are second- and third-year savings from PY8. Only in PY11, when the three-year EUL has expired, are new incremental annual savings claimed.

Table 30: Sample Compliance Calculations at Different EULs

Program Year	EDC #1 (1-year EUL) (kWh)	EDC #2 (3-year EUL) (kWh)
PY8	150	150
PY9	150	0
PY10	150	0
PY11	150	150
PY12	150	0
Phase III Total	**750**	**300**

EUL assumptions also impact TRC Test results. For example, EDC #2 would have a TRC ratio approximately three times that of EDC #1 in PY8, but a TRC ratio of 0.0 in PY9 and PY10. Like the calculation of savings for compliance, EDCs should perform Phase III TRC calculations consistently with the EUL approved in their Phase III EE&C plan.

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 listed on reports and some 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 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.

[106] The Phase II SWE HER persistence analysis addressed the concept of a savings decay rate. No Pennsylvania EDCs included a decay perspective in their Phase III EE&C plans so accounting nuances for this approach are not discussed in this protocol.

EDCs and their evaluation contractors may also consider focus groups with treatment households and businesses to learn more about their engagement with paper and electronic reports.

6.2 Demand Response Programs

6.2.1 Introduction

The Phase III Implementation Order for Act 129 EE&C programs established demand response performance targets for six of the seven EDCs subject to Act 129 and allowed for Penelec to voluntarily include a DR program. Table 31 provides an overview of the demand response initiatives approved in Phase III EE&C plans by the PUC.

Table 31: Summary of DR Offerings in Phase III EE&C Plans

Demand Response Initiative	PECO	PPL	Duquesne Light	West Penn	Met-Ed	Penn Power	Penelec
Residential DLC Switches	✓						
Residential Smart Thermostat	✓	✓					
Residential Behavioral DR	✓			✓	✓	✓	✓
C&I Load Curtailment	✓	✓	✓	✓	✓	✓	
C&I DLC Thermostats	✓						

While these offerings vary by delivery mechanism and targeted customer class, each initiative is a form of dispatchable, or event-based, conservation. It is important to distinguish the temporary impacts from these dispatchable programs from the “everyday” peak demand reductions produced by other EE&C programs. Unlike Phase I of Act 129, the Phase III demand reduction requirements are specific to demand response and cannot be satisfied with coincident demand reductions from energy efficiency measures.

The Phase III Implementation Order and subsequent Clarification Order provided clear instructions to EDCs about which hours would be used to measure DR performance (e.g., when to call DR events):

1) Curtailment events shall be limited to the months of June through September.
2) Curtailment events shall be called for the first six days in which the peak hour of PJM’s day-ahead forecast for the PJM RTO is greater than 96% of the PJM RTO

summer peak demand forecast for the months of June through September each year of the program.

3) Each curtailment event shall last four consecutive hours.
4) Each curtailment event shall be called such that it will occur during the day's forecasted peak hour(s) above 96% of PJM's RTO summer peak demand forecast.
5) Once six curtailment events have been called in a program year, the peak demand reduction program shall be suspended for that program year.
6) The reductions attributable to a four-consecutive-hour curtailment event will be based on the average MW reduction achieved during each hour of an event.
7) Compliance will be determined based on the average MW reductions achieved from events called in the last four years of the program.
8) The EDCs, in their plans, must demonstrate that the cost to acquire MWs from customers who participate in PJM's ELRP is no more than half the cost to acquire MWs from customers in the same rate class that are not participating in PJM's ELRP.

There were several important operational details that were not addressed explicitly in the Phase III Implementation Order or the Clarification Order. The SWE, TUS, and EDCs have discussed these issues collectively and reached consensus on the following clarifications.

- To support wholesale energy market operations, PJM provides an hourly load forecast online that is updated every 15 minutes. In mid-Summer 2018, PJM is eliminating the current location of its 7-day forecasts and moving it to Data Miner 2. EDCs should use the most recent 7-day forecast presented in Data Miner 2 at 10:10 AM as the forecast of record when determining whether the following day will be an Act 129 DR event or not.
- The 96% threshold and resulting Act 129 event dispatch determinations will rely solely on Table B-1 of the January PJM Load Forecast Report called for in the Phase III Clarification Order.
- Act 129 DR events are limited to non-holiday weekdays. This approach is consistent with PJM peak load criteria and the SWE's modeling assumptions in the Demand Response Market Potential Study.

Table 32 shows the Phase III DR goals by EDC for Phase III of Act 129. Compliance with these goals will be assessed by averaging the load reductions achieved in each DR event hour in PY9 to PY12.

Table 32: Phase III DR Goals by EDC

EDC	Phase III DR Target (MW)
Duquesne	42
Met-Ed	49
PECO	161
Penelec	0
Penn Power	17

PPL	92
West Penn Power	64

Based on the program design characteristics, there could be a variable number of event days and hours over the course of the four summers where the DR programs are active. It is important to note that the Phase III target is not an average of the four program-year averages. To calculate impacts in this manner would weight event performance in summers with fewer events more than event hours in summers with a larger number of events. A simple arithmetic average of all DR performance hours[107] in Phase III eliminates any weighting complications across events or years.

The Phase III Implementation Order included a second requirement designed to encourage performance across events and program years across the Phase. The Commission directed EDCs to obtain no less than 85% of the target in any one DR event. Event-specific estimates of load reduction are simply the mathematical average of the impacts observed during the four event hours.

One area of the flexibility in the prescribed DR program design is which four hours a DR event is called. Consider the following hypothetical example day-ahead forecast relative to PJM's summer peak demand forecast of 152,131 MW[108] for the 2016/2017 delivery year. Notice that only one hour (4 pm to 5 pm) in the day-ahead forecast for the RTO exceeds the program design threshold of 96% of PJM's summer peak demand forecast for the delivery year.

Table 33: Hypothetical RTO Combined Integrated Forecast Load (MW)

Date	Hour Beginning (EDT)	Hour Ending (EDT)	Day Ahead Forecast Load (MW)	% of 2016 Peak Demand Forecast
7/12/2016	12:00	13:00	137,983	90.7%
7/12/2016	13:00	14:00	141,178	92.8%
7/12/2016	14:00	15:00	144,068	94.7%
7/12/2016	15:00	16:00	145,285	95.5%
7/12/2016	16:00	17:00	146,806	96.5%
7/12/2016	17:00	18:00	143,155	94.1%
7/12/2016	18:00	19:00	140,417	92.3%
7/12/2016	19:00	20:00	136,462	89.7%

In this situation, an EDC would be required to dispatch the Act 129 DR program during hour ending 17:00, but would have some discretion regarding the start and end time of the event. Events need to last four contiguous hours, begin and end at the top of the hour, and be the

[107] As defined in the Implementation Order
[108] http://www.pjm.com/~/media/documents/reports/2016-load-report.ashx at page 52 (Table B-1)

same for all customers in an EDC service territory, so the specific options an EDC would have in this scenario are as follows:

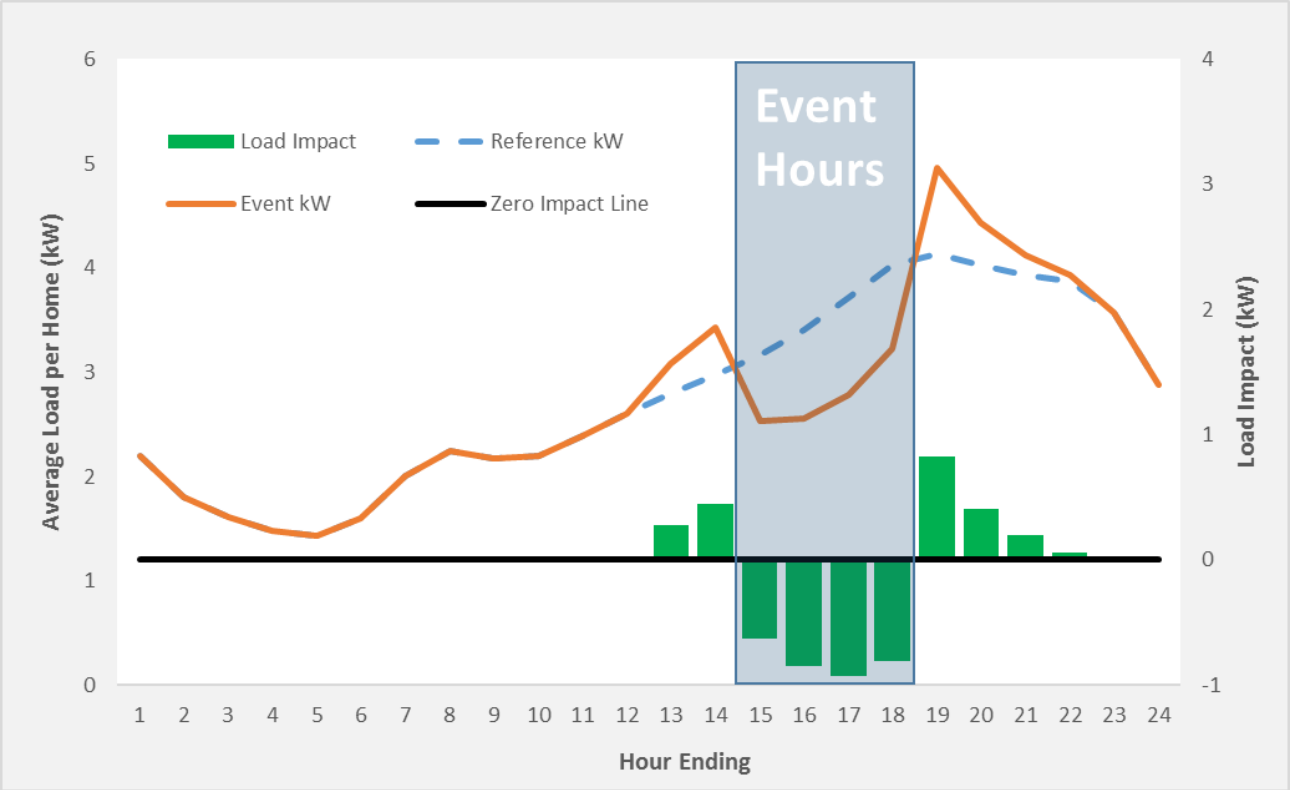
- Event begins at 1 pm and ends at 5 pm (HE14-HE17)
- Event begins at 2 pm and ends at 6 pm (HE15-HE18)
- Event begins at 3 pm and ends at 7 pm (HE16-HE19)
- Event begins at 4 pm and ends at 8 pm (HE17-HE20)

If there are multiple hours where the day-ahead forecast exceeds 96%, EDCs should select event hours to coincide with these hours. If there are more than four hours above 96% in the day-ahead forecast, EDCs have the discretion to select the four hours of the event. The selection of event hours along with the estimated hourly demand reductions should be included in EDC semiannual and annual reports to the PUC.

When calculating and reporting load impacts from Act 129 programs, it is important to remember that the compliance goals shown in Table 32 are at the system level. This means that impacts calculated at the retail meter need to be escalated by a line loss factor to account for transmission and distribution losses and calculate the reduction at the system level. Table 1-4 of the 2016 PA TRM provides the line loss factors by EDC and sector that are to be used when reporting Phase III DR performance.

The DR strategies the EDCs and their participating customers choose to employ to achieve demand reductions during event hours will often affect loads during the hours immediately before or after the four-hour event window. For example, an industrial plant may shut off a process at 1:30 pm in preparation for a DR event that begins at 2 pm. Conversely, homes or businesses may use extra electricity during the hours preceding or following event hours to minimize discomfort or make up for lost production. Another common example occurs with mass market programs that manipulate air conditioning usage. Figure 11 shows a hypothetical hourly impact graph for an AC load control program that exhibits load *increases* in the hours before and after the event window.

Figure 11: AC Load Control Example with Pre-Cooling and Snapback



The load increase in hours 13 and 14 of Figure 11 are a function of pre-cooling. Pre-cooling is a feature offered by many thermostat vendors where the device calls for more cooling than would typically occur in the hours preceding an event to minimize discomfort when cooling system operation is reduced during event hours. The load increases in hours 19-21 are often referred to as snapback and reflect the cooling system working harder than it would on a normal day to make up for reduced cooling usage during the event window. While this example deals with air conditioning usage, similar trends exist for other end-uses in both homes and businesses. It is important to note that load reductions or increases during surrounding hours in no way affect compliance with Act 129 DR targets. Although EDCs may wish to analyze loads in surrounding hours to accurately capture the energy (kWh) impacts of DR event calls for consumption reduction targets and the TRC test, compliance will be assessed exclusively on performance during event hours.

It is important to separate the treatment of load impacts in surrounding hours from load increases during event hours. Some participants may show metered loads above their estimated reference load during DR event hours. This positive impact (or negative load reduction) is factored into the mathematical average of event performance in the same manner as load reductions. Specifically, any zeroing out of load increases during event hours is not permitted. This requirement translates into EDC risk for Large C&I customers with erratic loads that should be considered during program enrollment. Section 6.2.2.1 provides additional guidance on how to deal with C&I customers with variable load patterns.

6.2.2 Gross Impact Evaluation

The objective of the impact evaluation is to estimate the verified gross peak demand (kW) impacts of the demand response program. The 2016 PA TRM includes two protocols that outline the core requirements for Act 129 demand response programs. This section of the Evaluation Framework is intended to provide additional technical guidance on the ex post evaluation protocols for Phase III DR offerings. While direct load control and behavior-based DR programs were addressed in a single protocol in the 2016 TRM, they are addressed separately in this document.

The focus of these protocols is ex post evaluations, or a retrospective analysis of the load impacts observed during actual DR events. EDCs may also gain important insights from ex ante evaluation, or the forecasting of future load impacts based on observed program performance. Although the data and methods used for ex ante forecasting of DR impacts are similar, this document does not provide specific guidance about ex ante evaluation other than noting which ex post methods are more useful for ex ante forecasting.

Section 5 of the PA TRM defines some key terms that are used frequently in this section. Those definitions are repeated below, along with some other key terms.

- **Observed Load** (*kW_Metered*): The actual measured electric demand in a participating premise, or group of premises, in a given period (usually an hour).
- **Reference Load** *(kW_Reference)***:** The counterfactual. An estimate of what electric demand would have been absent the DR program in a given period. The reference load is analogous to the baseline condition for an energy efficiency measure and sometimes referred to as the baseline or customer baseline.
- **Load Impact**: The difference between the observed load and the reference load in the period of interest in natural units (i.e., load reductions have a negative sign). Equal to kW_Metered – kW_Reference.
- **Load Reduction** (ΔkW): The difference between the observed load and the reference load in the period of interest with the sign flipped (i.e., load reductions have a positive sign). Equal to kW_Reference – kW_Metered.

6.2.2.1 C&I Load Curtailment

Load curtailment is a type of demand response where participants initiate actions within their facility to reduce loads in response to a notification from the EDC or CSP in exchange for financial compensation. Typically, the EDC does not have the ability to modify the operation of equipment within participant sites and relies on performance incentives, with or without the threat of financial penalties for non-performance, to produce load reductions during DR events.

The 2016 TRM established a hierarchy of methods for calculating gross verified savings for Act 129 load curtailment programs. It is important to distinguish the approach used to calculate gross verified demand reductions from the methods used to calculate settlements with individual customers. A 2013 LBNL report on M&V methods for demand response captured this distinction clearly and succinctly. "*More accurate program-level results can*

typically be obtained by using impact estimation methods that are not practical for settlement applications."[109] Key differences include the following:

- Customer settlements based on performance necessitate a separate estimate of the load reduction delivered by each program participant. Act 129 compliance goals are established at the EDC level so that pooled analysis methods are possible.
- Transactions in wholesale markets like those operated by PJM need to clear and settle quickly to function. Act 129 demand response programs have a less aggressive schedule. This allows for more complex models and data from longer time horizons to be used in the ex post evaluation.

Based on these considerations, the Commission established a preference for comparison group and regression methods over the "high X of Y"-style customer baseline (CBL) approaches favored by wholesale markets like PJM for customer settlement. The results of several independent studies[110] have shown that CBL methods produce reliable impact estimates when "high X of Y" day-matching is used in combination with a same-day symmetric additive or multiplicative adjustment to calibrate the comparison days to the observed loads on the event day prior to dispatch. However, the decision to use day-ahead event notification for Act 129 DR events in Phase III makes the use of same-day adjustment problematic because the event notification will likely cause participants to modify electric consumption during the hours prior to the event. This could take the form of pre-cooling, altered scheduling of processes, or any number of other factors. While these load impacts in surrounding hours do not directly affect performance, including them in the mathematical calculation of the reference load could distort estimates of what load would have been absent the DR event and bias load reduction estimates.

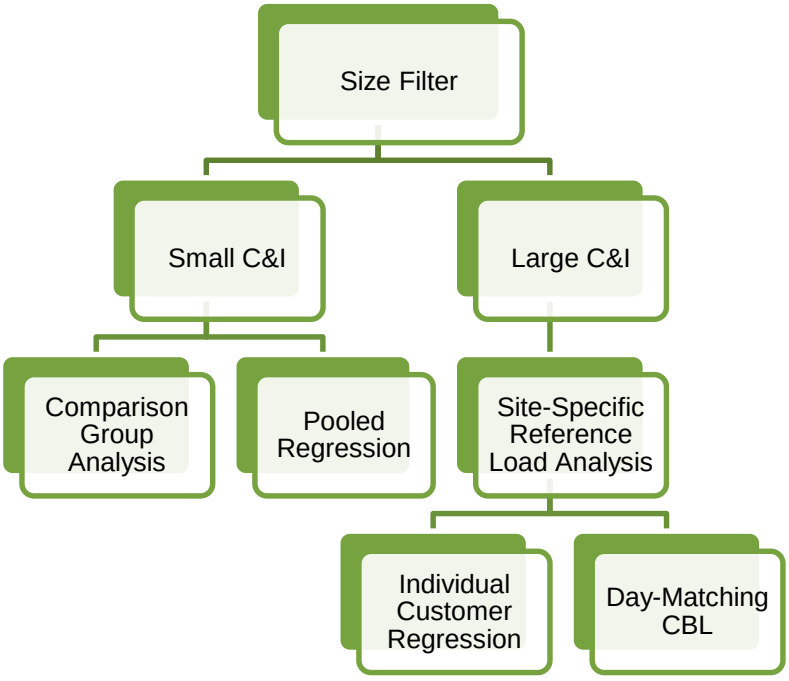
For operational simplicity, EDCs may choose to contract with CSPs and pay customer incentives based on CBL methods because they are more transparent and easier to calculate and track in real-time. Following the TRM hierarchy of methods for load curtailment can create uncertainty for these EDCs because there is a possibility that the more rigorous methods used to calculate gross verified savings will return higher or lower impact estimates than the CBLs and create misalignment between incentive payments and compliance savings. This risk can be mitigated to some extent by careful selection of the CBLs used by the CSP to calculate customer settlements.

During Phase I of Act 129, EDC load curtailment programs were very top-heavy, with a small number of very large sites contributing the majority of the load reduction for the program. The SWE encourages evaluation contractors to allocate evaluation resources in a similar fashion and focus efforts on producing accurate and defensible reference loads for the large sites that are delivering the majority of compliance savings. A hypothetical process is presented in Figure 12. For the smaller sites, comparison group analysis and pooled regressions allow evaluation contractors to quickly estimate gross verified savings

[109] *Measurement and Verification for Demand Response*. Prepared for the National Forum on the National Action Plan on Demand Response. https://emp.lbl.gov/sites/all/files/napdr-measurement-and-verification_0.pdf
[110] http://www.ieso.ca/Documents/consult/drwg/drwg-20140603-Item5b%20Final.pdf

for a large number of sites. The SWE may also approve an approach where evaluation contractors select a sample of small sites from the program population and adjust CSP settlement calculations via a realization rate. For large sites, the reference load calculation method should be studied carefully for each site and selected on the basis of accuracy, precision, and bias.

Figure 12: Sample Load Curtailment Evaluation Process



The split between "small" and "large" sites should be clarified in the EM&V plan for the program. Possible strategies include the following:

- Rate class or sector (Small C&I vs. Large C&I)
- Peak load contribution (e.g., < 500 kW = small, ≥ 500 kW = large)
- Expected peak demand reduction (on a percent or absolute basis)
- Business type (e.g., Commercial = small, Industrial = large)
- Minimum acceptable match quality (discussed in Section 6.2.2.1.1)

The following sections provide technical guidance on applying the load curtailment methodologies outlined in the 2016 TRM. The metering protocols assume that EDCs have hourly or sub-hourly meters (interval metering) for all participating sites and a large share of non-participating accounts. Many PJM DR participants utilize pulses obtained from EDC meter devices and CSP pulse counters/Energy Management systems to track and manage DR performance in real-time, as well as to report DR event results to PJM. The EDC meter devices generate pulses at a rate proportional to the energy consumed. These pulses can then be recorded by CSP pulse counters/Energy Management Systems and converted to average electric demand data for analysis and reporting comparable to interval data from

EDC revenue meters. Pulse meter data can be used as a substitute for interval data from the EDC revenue meter provided certain conditions are met.

- Pulse meter data reflect EDC metered premise load and not just load from a sub-panel or specific process within the participating facility.
- The EDC or EDC evaluation contractor performs an analysis on a sample of sites where pulse meter data are proposed as a surrogate for revenue meter data to corroborate the accuracy and consistent availability of the pulse meter data. This validation exercise should be described in the EM&V plan for the program.

6.2.2.1.1 Matching

A true experimental design such as a randomized control trial (RCT) is generally not practical for load curtailment programs. DR events, by nature, require interruptions to normal business operations, so EDCs operate programs on an "opt-in" basis rather than defaulting customers into the load curtailment program. Similarly, holding back willing DR participants to serve as a control group creates challenges for goal attainment and equity across the rate class. In the absence of a randomized evaluation design, comparing loads of participating businesses with loads of non-participating businesses can provide a reasonable estimate of the counterfactual—or what the loads of DR participants would have been absent the DR event call. Weather conditions and other day-specific factors are controlled for because non-participants and participants experience identical weather conditions and observable externalities such as whether local schools are in session or if there is a home baseball game. The problem with this approach is selection bias. The non-treatment of selection bias is illustrated[111] in Equation 19 and the discussion that follows.

Equation 19: Comparing Outcomes Across Participants and Non-Participants

$$kW_i = \alpha * X_i + \beta * T_i + \varepsilon_i$$

Where:

kW_i = Electric demand of participant i

X_i = Array of observed characteristics about participant i

T_i = An indicator variable equal to 1 for participating businesses and 0 for non-participants

ε_i = An error term containing unobserved characteristics that affect the kW term

α and β = Regression coefficients

Estimating Equation 19 in a regression framework will return an estimate of the DR program effect (β), but the coefficient may contain bias for either of the following reasons:

- The DR program was marketed to or targeted certain customers because of some characteristic or set of characteristics not reflected in the non-participant control

[111] Example adapted from Khandker, et al. (*Handbook on Impact Evaluation: Quantitative Methods and Practices*. World Bank Publications, 2010)

group such as size, business type, location on distribution system, prior DR participation, etc.

- Certain types of businesses self-select into the DR program because it makes business sense for their facility. For example, agricultural operations that have large pumping loads to move water for irrigation have a lot of flexibility in terms of when energy intense motor operation occurs (e.g., crops require a lot of water, but are not particularly sensitive to timing, so motors can be turned off for a few hours for a DR event without disrupting operations).

This leads to a case in which the treatment variable (*T*) is correlated with observed characteristics (*X*) and the error term (ε). A fundamental assumption of ordinary least squares (OLS) regression is that each explanatory variable is uncorrelated with the error term. The correlation between *T* and ε is referred to as endogeneity and potentially biases other estimates in the equation, including the parameter of interest, β.

Propensity score matching (PSM) is one matching technique that EDC evaluation contractors may use to minimize the bias introduced by comparing participants to non-participants. The premise is to select a subset of the non-participating businesses within the EDC service territory that are most similar to the participating businesses across observable characteristics. Forcing similarity across observable characteristics controls for differences in observables directly and reduces the likelihood that businesses will differ on unobservable dimensions, thus reducing the threat of selection bias.

The first step of PSM is to create a statistical model of program participation. The output of a binary outcome model is a value between zero and one for each account that represents the home or business's likelihood of participating in the program. This probability, or propensity score, is then used to match non-participants with participants who share similar characteristics. Equation 20 provides formal notation for a logistic regression model. The log of the odds[112] of an outcome (y) is modeled as a function of one or more explanatory variables (x).

Equation 20: Logistic Regression Notation

$$\log\left(\frac{p(y)}{1-p(y)}\right) = \beta_0 + \beta_1 * X_1$$

Binary outcome models, such as logistic regression and probit regression, allow researchers to model processes that have only two possible outcomes. In this case, EDC accounts that are enrolled in the DR program are coded as 1, and accounts that are eligible but not enrolled are coded as 0. Table 34 shows an abridged version of the type of data set evaluation contractors might construct for matching purposes. Notice that the data set is structured in a "wide" format with a single row for each site. The "participant" field denotes whether or not the site is enrolled in the DR program. The other columns in the table contain relevant observable characteristics about the sites.

[112] Odds is equal to probability of success (coded as a 1) divided by probability of failure (coded as a 0).

Table 34: Data Structure for Binary Outcome Model

Site ID	Participant	Business Type	Annual kWh	mean15	mean16	mean17	mean18	Load Factor
1	1	Retail	3,708,167	521	525	518	504	0.66
2	1	Retail	2,957,093	425	422	415	406	0.66
3	1	Office	4,246,764	605	607	608	591	0.68
4	1	Office	2,835,827	373	375	370	363	0.71
5	1	Office	4,130,552	534	526	530	522	0.77
6	1	Education	4,468,862	605	603	603	606	0.70
7	1	Education	5,647,900	764	766	754	742	0.67
8	0	Retail	2,321,976	323	325	316	310	0.64
9	0	Retail	3,872,764	551	547	554	548	0.67
10	0	Retail	3,117,763	400	393	396	388	0.79
11	0	Retail	3,594,885	499	505	504	489	0.62
12	0	Retail	1,442,605	197	197	203	202	0.68
13	0	Office	4,175,250	567	572	568	578	0.72
14	0	Office	5,004,513	718	731	726	719	0.65
15	0	Office	4,574,169	658	643	659	648	0.64
16	0	Office	5,075,656	769	758	775	764	0.60
17	0	Office	5,810,178	831	840	848	823	0.63
18	0	Office	5,127,520	727	725	717	714	0.69
19	0	Office	4,296,673	607	608	601	586	0.65
20	0	Office	4,685,683	668	675	678	674	0.64
21	0	Education	4,607,431	667	676	673	656	0.65
22	0	Education	4,469,159	639	646	641	630	0.65
23	0	Education	4,148,298	578	586	582	568	0.66
24	0	Education	4,197,209	632	631	627	612	0.63
25	0	Education	4,077,648	580	585	584	564	0.66
26	0	Education	3,640,150	534	526	530	523	0.61
27	0	Education	4,108,079	559	557	558	542	0.71

The fit of the binary outcome model and the subsequent quality of matches may improve if evaluation contractors are able to add more descriptive variables and precisely estimate their effects. Evaluation contractors may consider a multi-step procedure for selecting the covariates of a propensity score model that uses stepwise selection and allows the for candidate variables to enter on a standalone basis, as interactions, and as squares.[113] One challenge with implementing this approach is that the data must be available for both participants and non-participants to be useful—and non-participant data are frequently limited. Electricity consumption is obviously the most important observable characteristic. Although not ideal, it is possible to build quality matches with only load data. There is a number of load characteristics that evaluation contractors may wish to consider. Table 34

[113] Imbens and Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences.* Page 342

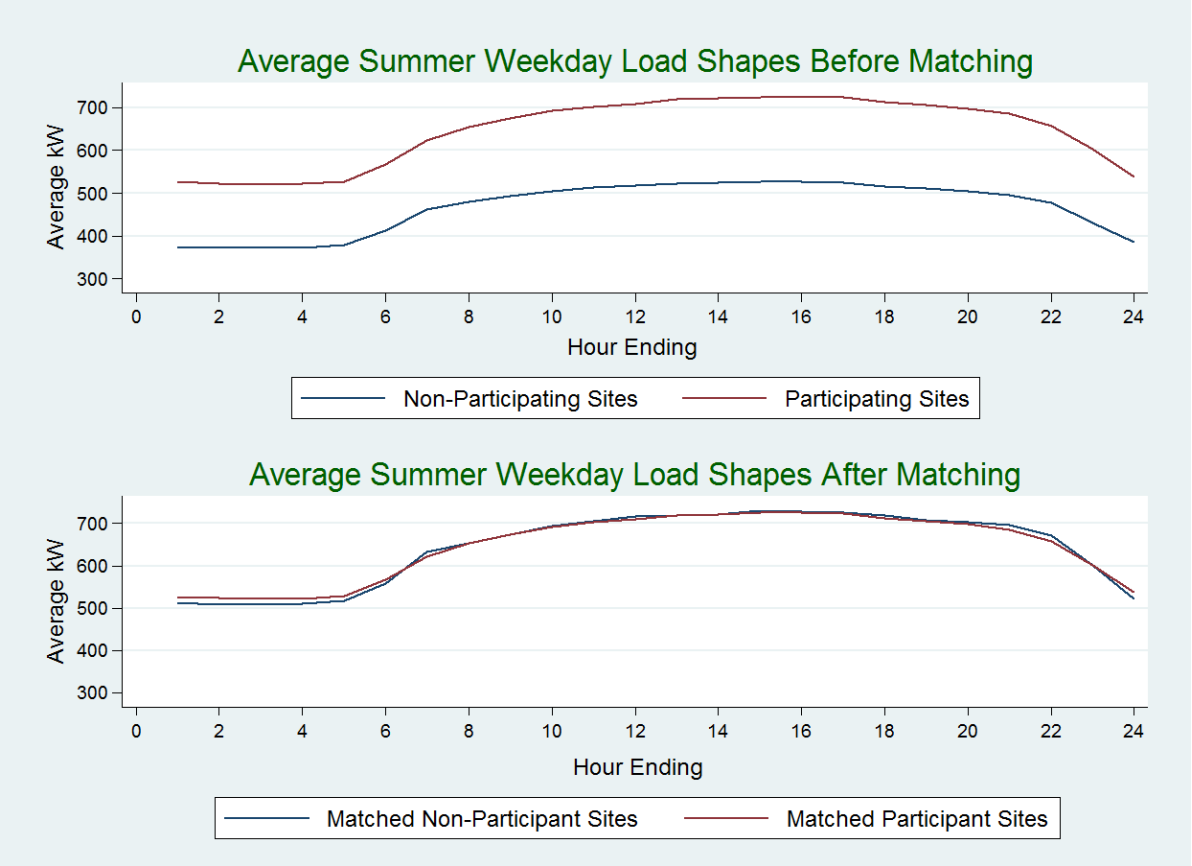
includes an array of columns named mean15 to mean18. These variables represent the average hourly loads of the sites on summer weekdays in hour ending 15, 16, 17, and 18 (i.e., when DR events are likely to occur). Other descriptive statistics evaluation contractors may wish to consider whether binary outcome models include the following:

- **Peak load contribution** (PLC) – Calculated based on metered loads during system peak hours. Some EDCs or EGSs use this metric to allocate capacity cost to customers.
- **Slope of load duration curve** – The coefficient of a simple regression line fit through a year of hourly loads sorted in descending order.
- **Load factor** – The relationship between average usage and peak consumption. Calculated as either mean/max or max/mean.
- **Weather sensitivity** – How sensitive is a home's or business's electric consumption to changes in weather conditions? It is often useful to calculate separate correlation coefficients for the heating season and cooling season.

Evaluation contractors should filter out accounts with PJM registrations from the Act 129 non-participant pool to avoid selecting accounts who curtail load for PJM on an Act 129 event day because this would bias the reference load downward and unfairly reduce EDC load reduction estimates.

Figure 13 shows output from a simple matching exercise. Notice that before the binary outcome was estimated and the output was used to select similar accounts, the average DR participant site used much more electricity, on average, than the pool of non-participants. After estimating a logistic regression model, selecting the non-participant(s) with the nearest propensity score to each participant, and discarding the unmatched non-participants, the load shapes of the two groups show much better alignment.

Figure 13: Improved Alignment via Matching



It is worth noting that the y-axes of the load shapes shown in Figure 13 each reflect an *average* across multiple sites. While evaluation contractors could certainly find a match or matches for each DR participant and aggregate impacts calculated at the site level, a pooled approach smooths out some of the noise of individual customer variations in consumption. A pooled approach also leverages the fact that EDCs do not need customer-specific estimates for Act 129 compliance reporting. With 1:1 matching, a sum across sites would achieve the same function and require one fewer step to calculate program-level savings.

Once a binary outcome model has been developed and estimated, evaluation contractors need to actually match each participant to one or more non-participants based on the propensity score. The example described above and illustrated in Figure 13 uses a simple technique referred to as nearest neighbor matching, where the smallest absolute value of the difference in propensity score is used as the matching criterion. There are other matching techniques that can be used to select matches from binary outcome models. EDC evaluation contractors are encouraged to consider different approaches and select their preferred method based on their professional judgment.

This protocol does not recommend one matching technique over another. As stated in the 2016 TRM, evaluation contractors are to choose the technique used to select the comparison group based on their professional judgment. The key metric when selecting a

binary outcome model specification and matching technique is how well the comparison group's loads compare to the participant group's loads in out-of-sample testing. The general framework for out-of-sample testing is as follows:

1) Identify a small number of event-like days when DR events were not called. Hot summer weekdays during the summer of 2016 will be ideal.
2) Remove these days from the primary "training" data set and save in a "validation" data set.
3) Select the comparison group via various matching methods.
4) Compare electric loads of the DR participants to the selected comparison group on the validation days. Pay special attention to likely event hours (afternoons).
5) Compute metrics of bias, accuracy, and precision.
6) Select the matching method and resulting comparison group based on performance across these metrics.

Matching can be performed either with or without replacement. Matching with replacement means that a given non-participant can be matched with more than one DR program participant. Matching without replacement means that once a non-participant is selected as a match, they are removed from the eligible pool. Either approach is acceptable for Act 129 DR program evaluation. The "with replacement" decision is not likely to be terribly meaningful, especially for residential programs where the non-participant pool includes hundreds of thousands of households. Matching without replacement will necessarily reduce the quality of certain matches but makes post-estimation simpler because each site has the same weight in the sample. When matching with replacement is used, sample weights need to be calculated and applied. Stated simply, if a non-participant site is matched to three different participant sites, it needs to carry three times the weight in the reference load calculation as a site that was matched to just one participant.

In Phase I of Act 129, load curtailment programs attracted some of the largest and most unique accounts in EDC customer bases. If Phase III load curtailment program participation is similar, it is expected that for most of these accounts, there simply will not be a non-participating customer within the EDC service territory who uses electricity in a similar enough way to provide a suitable match. Perhaps there are four steel mills in an EDC service territory, all four have enrolled in the program, and no other accounts in the service territory exhibit the same consumption levels or load volatility. The concept of a caliper is useful to identify such participants who do not have suitable matches. A caliper is typically the maximum tolerable difference in propensity score between a participant and their match. Evaluation contractors may also choose to place a caliper on specific key variables either with or without a caliper on propensity score. For example, an evaluation contractor may choose to require no more than a 10% difference in peak load contribution for a match to be considered successful.

The third approach in the TRM hierarchy of methods for load curtailment programs suggests a hybrid regression-matching approach that works well with a caliper-based matching approach. Specifically, the TRM refers to "*a hybrid Regression-Matching method where matching is used for most customers and regression methods are used to predict reference loads for any large customers who are too unique to have a good matching*

candidate." One practical application of this protocol would be to run all DR participants through a matching exercise and evaluate the successful matches via a comparison of loads with the selected matched control group. Participants who cannot be matched successfully would need to have load impacts evaluated using an alternative method such as an individual customer regression or a day-matching CBL.

Once a comparison group has been selected and constructed for a load curtailment participant or group of participants, evaluation contractors still need to estimate the load impact associated with each DR event hour by comparing the observed loads of participants and the selected comparison sites. If the quality of the matches is excellent and the groups show essentially no differences in load patterns during non-event hours, this calculation is a simple difference in means (or sums). In practice, there will likely be minor differences between the load shapes of the two groups. Evaluation contractors should consider applying difference-in-differences techniques to correct for these small differences in consumption. Such corrections can be implemented via regression or a more manual process where the average difference between group loads on non-event weekdays during the hour of interest are calculated and then subtracted from the observed difference during the DR event hour of interest.

6.2.2.1.2 Regression

Regression analysis is a calculation method that estimates a mathematical relationship between a dependent variable (measured electric load) and other variables that help explain the observed variability in loads. Regression is a broad category that includes a number of estimation algorithms, functional specifications, and econometric correction techniques to deal with various issues. Because of the heterogeneous nature of C&I customers and their load patterns, no single technique will work for all sites. This protocol is intended to provide high-level guidance, but ultimately the statistical training of EDC evaluation contractors will govern the selection of the most appropriate regression method for a given application.

Regression analysis is especially useful when ex ante forecasts of load reductions are desired. This protocol is focused on ex post analysis, but the regression methods can be used with little modification to estimate program capability under a range of conditions by interacting event indicator variables with key independent variables so that expected loads are expressed as a function of some condition. The relationship between weather conditions and load impact is of key importance for many DR customers. Equation 21 presents a simple model that expresses electric load as a function of the hour of the day and outdoor air temperature.

Equation 21: Sample Weather-Dependent Regression

$$\mathrm{kW_t} = \sum_{\mathrm{h}=1}^{24} * I_h * \beta_{\mathrm{h}} \; + \; \gamma_h * \sum_{\mathrm{h}=1}^{24} * I_h * \mathrm{AirTemp_t} \; + \; \varepsilon_{\mathrm{t}}$$

Table 35 defines the model terms and coefficients in Equation 21.

Table 35: Sample Weather-Dependent Regression Model Definition of Terms

Variable	Definition
$\mathbf{kW_t}$	Metered electric demand in time period t
$\mathbf{I_h}$	An indicator variable that equals one during hour h, and zero otherwise. This variable models each hour's difference from the reference hour
$\boldsymbol{\beta_h}$	The coefficient on the standalone hour indicator variable. The model intercept for hour h
$\mathbf{AirTemp_t}$	Dry bulb temperature (F) in time period t taken from some reliable meteorological source
$\boldsymbol{\gamma_h}$	The coefficient on the hourly indicator variable interacted with outdoor air temperature. Represents the expected change in load (kW) for a one-degree (F) increase in outdoor air temperature
$\boldsymbol{\varepsilon_t}$	The error term

If Equation 21 were utilized for a DR customer or group of customers, evaluation contractors would use the β and γ coefficients along with hourly weather records for the DR event day of interest to estimate the reference load. The weather variable in Equation 21 is the dry bulb temperature expressed as degrees Fahrenheit. Temperature could also be expressed on a Celsius or Kelvin scale, if desired. This is just one of many weather variables evaluation contractors might choose to test in a regression model. Others include the following:

- Wet bulb temperature
- Temperature Humidity Index (THI)
- Wind speed
- Dew point
- Precipitation

Transformations of weather variables may be used at the discretion of evaluation contractors. For example, cooling degree hours (CDH) is a shift of the base temperature scale where some base temperature is subtracted from the measure condition. For example, if the outdoor air temperature is 95 degrees (F) in a given hour, this could be expressed as 35 CDH on a base 60 scale. This transformation will produce identical predictions to the native scale for temperatures above 60 degrees (F), but with a more intuitive intercept term. The model's ability to predict load at 0 degrees (F) is irrelevant because temperatures would never approach zero during a summer DR event; however, it can be confusing to see intercept values that represent expected load at 0 degrees (F). When a CDH transformation is used, the model intercept becomes a proxy for expected load absent air conditioning, which is a more tangible concept for stakeholders.

The model specification shown in Equation 21 only considers air temperature during the hour of interest. Oftentimes, inclusion of both current weather conditions and conditions leading up to the hour of interest will improve model fit because it better addresses thermal inertia considerations of the HVAC system within the building. Inclusion of multiple weather variables can be beneficial for certain models, but evaluators should be mindful of including

terms that are highly correlated and thus carry too much duplicate information because this can lead to multicollinearity and inconsistent estimates. Variance inflation factor (VIF) statistics are a useful tool to assess the degree of multicollinearity among independent variables and is available in most statistical packages.

The example model specification shown in Equation 21 includes indicator variables for the hour of the day. This allows the model to fit the daily time trends within a facility and predict load specifics to a given hour of the day using observed load data from that hour. Indicator variables for the day of the week or day type (weekday vs. weekend/holiday) can greatly improve the predictive ability of models. Consider a standard office building that is mostly occupied Monday through Friday and experiences all of the lighting, computer equipment, cooling, and ventilation loads typical of an office setting. That same facility is mostly empty on Saturdays and Sundays with most lights turned off and cooling and ventilation settings relaxed. Even with identical weather conditions, electric loads would be expected to be much lower on weekends than weekdays. Since the goal is to produce the best possible estimate of what loads would have been absent the DR dispatch, basing reference load estimates for a weekday event on observed non-event weekday loads would be advantageous.

An alternative strategy to indicator variables for the hour of the day or day of the week is to run multiple models—one for each level of the variable. The downside to either approach is that, by segmenting the data into more specific bins, the number of observations the regression model has to estimate coefficients is reduced. Consider an approach where indicator variables (or separate models) are used for both day of week indicators and hour. Now assume that a DR event is called on a Wednesday afternoon in August and the reference load for HE17 (4:00 to 5:00 pm) is to be estimated. The highly segmented model means that there are fewer than 20 measurements that summer to inform the estimate because the model will limit itself to only other readings from HE17 on Wednesdays.

When a regression-based approach is used, evaluation contractors should ensure that the model has a sufficient number of observations. Practically speaking, this means including a large number of non-event days in the model. As the number of indicator variables or separate models increases, this become especially important. Evaluation contractors are encouraged to consider the following approaches to improve the explanatory power of models:

- Both historic days and future days. If a DR event is called in June, data from July to September can be used in the regression.
- Days from outside of the Act 129 DR season (June to September)
- Days from the previous summer

With certain customers, there is a tendency for days temporally close to one another to be more similar than days that are farther apart, all other things being equal. Consider a food processing plant whose production ramps up toward the end of growing season and then gradually slows down once the harvest is processed. When data from a long time horizon are used in the regression model, evaluation contractors may choose to explore corrections that deal with the varying lengths of time between non-event days and the event day of

interest. One approach is to use a weighting scheme that places greater importance on the days close to the event than days more distant from the event. Another option is to utilize a time-series model that explicitly addresses the autocorrelation. Autoregressive integrated moving average (ARIMA) models and Prais-Winsten estimation are two time-series options that specifically address autocorrelation in functional form and may prove useful for customers where loading patterns change over time. Evaluators need to be especially mindful of excluding hours between event notification and event start from these models.

Weather variables are great predictors for many customers, but other accounts show little or no weather sensitivity. This is especially true for the largest industrial customers/accounts that also tend to be among the largest providers of DR in Pennsylvania. Selecting explanatory variables for these accounts can be challenging because loads are driven almost exclusively by business processes that are invisible to the evaluation contractor. One variable that can prove useful for these Large C&I accounts is the wholesale price of electricity. Unlike most Residential and Small C&I accounts, who pay a flat rate per kilowatt hour consumed, most Large C&I customers pay electricity rates that vary hourly in order to mirror real-time conditions in the wholesale electric market.

Wholesale electric prices generally vary according to fluctuations in supply and demand. Demand will often push prices higher during peak times like hot summer afternoons and generally remain lower during periods when demand on the system is lower. By including the Locational Marginal Price[114] (LMP) of electricity as an independent variable in the regression model, evaluators can capture tendencies of customers to reduce cost by dodging high-priced energy and running energy-intensive processes when it is less expensive to do so. This behavior is a form of demand response in itself but is completely independent of Act 129 and therefore should be reflected in reference load calculations.

Occasionally, evaluation contractors will encounter load curtailment participants whose loads are highly variable in a way that is not explained by the typical explanatory variables discussed previously. Depending on enrollment terms, these accounts can constitute a risky investment for EDCs because they tend to produce erratic load impact estimates. Negative load reduction estimates that are just noise can pull down program averages for a given event even though it will usually even out over multiple events. If such customers are enrolled and evaluators are unable to produce reliable reference load estimates, a possibility is to gather supplemental information from the participant. LBNL's M&V Working Group Guide to Demand Response recommends that program administrators "*allow a customized baseline that uses additional operational information supplied by the participant*."[115]

If this approach is implemented, evaluation contractors need to be very cautious to ensure that the supplemental data are useful but not biased by the DR event itself. Consider the example of a concrete aggregate plant that uses large machinery to crush rock into smaller

[114] Historic real-time and day-ahead LMPs by zone can be downloaded from PJM's website at http://www.pjm.com/markets-and-operations/energy/real-time/monthlylmp.aspx

[115] *Measurement and Verification for Demand Response*, prepared for the National Forum on the National Action Plan on Demand Response, p. 38 (PDF p. 66).

pieces for various applications. These facilities have erratic loading patterns that make reference load estimation challenging. The hourly run time or operating schedule of the rock-crushing machine(s) would have tremendous explanatory power for interval facility loads, but causes issues because this is precisely the strategy the plant would use to shed load during a DR event. If a regression model included hourly machine run time as an independent variable, the reduced machine run time during the event would cause the model to predict *event* load rather than reference load. One solution is to gather scheduled production or weekly production totals instead. So in this concrete plant example, it would be assumed that the DR event did not change the business's net output requirements (e.g., total number of tons of rock crushed or truckloads shipped for the week), only what occurred during the event hours, and that this output was subsequently made up for later. The use of supplemental data should be considered a last resort and applied only when evaluators have a good understanding of the affected business processes within the facility and have exhausted other strategies to develop an accurate reference load.

As discussed in Section 6.2.2.1.1, gross verified DR impacts for Act 129 are assessed at the program level. While customer-specific impact estimates may be needed for settlement, evaluation contractors may make use of pooled regression models. Pooled models tend to be less noisy than individual customer models and can be especially useful for accounts with more than one location in the program.

6.2.2.1.3 Day Matching

The 2016 PA TRM includes the Commission's strong encouragement that, where possible, EDCs utilize comparison group analyses, within-subject regression analyses, or hybrid regression-matching instead of day-matching approaches, also referred to as CBL or "high x of y," for calculating gross verified impacts from C&I load curtailment programs. However, the TRM did not prohibit the use of CBLs and even noted conditions when they tend to produce valid results. The underlying estimation method used in CBL methods is averaging. This means the reference load calculation is the simple arithmetic mean of loads from the same hour on non-event days. Table 36 provides an example using a common CBL method used for PJM settlement. In this example, the second Friday has a DR event from 1:00 to 5:00 pm. This version of the "high 4 of 5" method treats weekdays as a single day type, so it looks back over the last five weekdays and selects the four days with the highest loads during the event hours. Wednesday has the lowest loads of the five days, so it is excluded. Hourly loads from the other four days are averaged to calculate the reference load for each hour.

Table 36: High 4 of 5 CBL Calculation

Hour Ending	Friday	Monday	Tuesday	Wednesday	Thursday	CBL for Friday Event
14	250	220	290	~~175~~	240	*250*
15	275	265	295	~~190~~	260	*274*
16	300	305	320	~~180~~	280	*301*
17	310	350	270	~~160~~	30	*240*
4-Hour Total	**1135**	**1140**	**1175**	**~~705~~**	**810**	***1065***

Although less accurate than more rigorous methods like matching and regression, this approach is simple and transparent, which makes it a great operational choice for markets that need to clear and settle quickly and CSPs who need to track real-time performance. The result in the rightmost column of Table 36 is actually identical to what would be estimated by the regression model shown in Equation 22.

Equation 22: Day Averaging via Regression

$$\mathrm{kW_t} = \sum_{\mathrm{h}=1}^{24} * I_h * \beta_{\mathrm{h}} + \varepsilon_{\mathrm{t}}$$

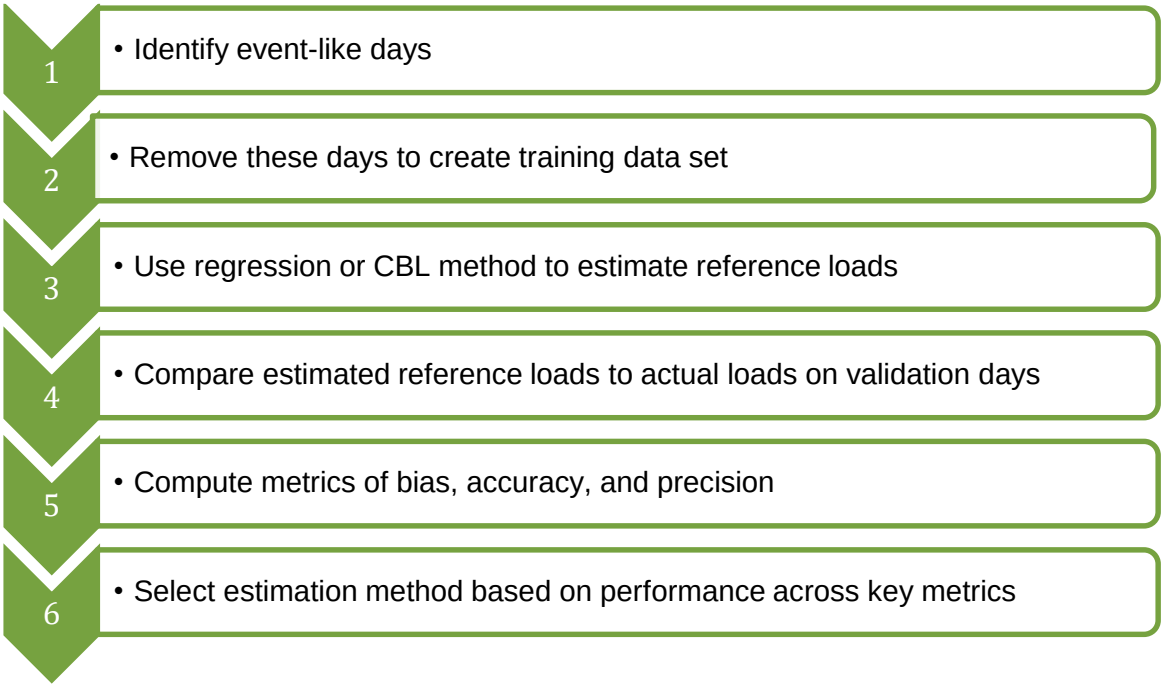
Equation 22 is essentially averaging via regression, or a regression model without any explanatory variables (R^2=0.00). The implication of this comparison is that a regression model without any meaningful predictor variables (weather, LMP) will not offer any improvement over a CBL because the calculation is essentially the same. Therefore, for facilities with constant loads that are not weather sensitive, CBL methods are a reasonable option.

A “high x of y” calculation like the one shown in Table 36 clearly places an emphasis on recency. Selected days within the lookback are included in the calculation, and each is weighted identically. Days outside of the window are excluded entirely. By design, most CBL methods are a very rigid set of calculation rules compared to regression analysis. However, it is possible for evaluation contractors to incorporate additional flexibility into day matching approaches. Some useful techniques include the following:

- Assess the right x and y independently by participant. One overarching theme in this section is that one-size-fits-all approaches will reduce the reliability of ex post load impact estimates. Explore many possible reference load calculation methods and select the best fit for each customer.
- Consider using days after the event in addition to days prior to the event. This allows for higher y’s and leverages the length of time available for Act 129 evaluations.
- Explore weighting of days. Traditional CBL methods have two weights—in or out. If loading patterns exhibit temporal trends, evaluators may consider weighting days by proximity or even by day of the week (e.g., Mondays and Wednesdays carry more weight than Fridays in the reference load calculation for a Tuesday.)

Figure 14 shows the general steps that evaluation contractors should take to select the reference load calculation for a given participant or group of participants.

Figure 14: Reference Load Selection Steps



6.2.2.1.4 On-Site Generation

Some participants in EDC load curtailment programs may use on-site generation to achieve a reduction in the amount of energy they take from the grid. This curtailment strategy is acceptable for Act 129 DR programs, but participants need to be mindful of environmental regulations that limit the number of hours per year that certain types of machines are allowed to run for economic purposes. The reference load calculation methodologies described previously can be used to estimate load impacts for participants who use on-site generation, but evaluation contractors may also choose to leverage the metered output of the generator as a proxy for load reduction. This approach makes several assumptions that should be validated by the EDC/CSP before accepting metered output as gross verified demand reduction. Key assumptions include the following:

1) The self-generated power is used in full by the facility or connected facilities during the DR event and not stored or used to power some atypical process.
2) On-site generation is limited to DR events and not used as part of normal operations.

6.2.2.1.5 Exclusion of Other Event Days

The intent of reference load calculations for load curtailment participants is to develop an estimate of what loads would have been in the facility if no DR events were called. It follows that other event days should be excluded from regression or day-matching calculations so that reduced loading conditions from other DR events do not cause load reduction estimates to be understated. This process is relatively straightforward for other Act 129 demand response events because evaluation contractors will know exactly which dates and

hours Act 129 events were called, and these dates and hours will be identical for all participants in the program.

For DR participants who also participate in PJM's Emergency (capacity) or Economic (energy) markets, removal of other events can be more complicated. Ultimately, failure to remove PJM events from reference load calculations will harm the participating site (by reducing their compensation) and EDC (by understating program performance). Open sharing of PJM curtailment details is in the financial interest of both parties, so this protocol assumes parties will overcome any data exchange barriers and that evaluation contractors will have this information for all accounts. If PJM event details are not known via interchange accounting records or some other reliable method, they cannot be removed. Evaluation contractors should never review load shapes and guess which days the account curtailed load for PJM. Adding back PJM load reductions to metered loads to reconstruct a baseline day is also prohibited.

It is possible that customers will curtail load for both Act 129 and PJM on the same day. When this happens, the start and end times of the two curtailment periods will often differ. The prohibition of day-of adjustments in the 2016 TRM simplifies this issue considerably because the start and end times of the PJM event are irrelevant to Act 129 peak demand impact calculations. If EDCs and their evaluation contractors are interested in the energy impacts of Act 129 event calls, calculations of pre-event changes should end when the earlier of the two events begins. Similarly, post-event analysis should begin when both events have concluded.

Evaluation contractors may encounter a situation where an Act 129 participant participates in the PJM energy market virtually every day. This creates challenges for reference load development because there are very few non-event days to use in the calculation. It can also lead to a "frozen baseline," where a site's reference load never gets updated because every day is an event day. Evaluation contractors should consult the SWE if there are large sites that have energy bids clear every day. Possible mitigation strategies include using weekends in the reference load calculation or establishing a conservative peak load contribution for the site to use as the reference load for Act 129 load impact estimates.

Some C&I facilities schedule plant closures for maintenance at various times during the year. This can result in several days of abnormally low loads because of dramatically curtailed operations. To the extent that such closures are known and documented, evaluation contractors may treat these days similar to event days and exclude them from the reference load calculation. If an Act 129 DR event coincides with a plant closure, EDCs and their evaluation contractors can use the normal reference load methodology (excluding other known closure days). In the scenario, the facility has already reduced its load, so it has no other way to reduce consumption further and achieve committed load reductions. The fact that the load reduction would have happened absent the Act 129 program is an attribution issue and should not impact the gross verified savings calculations used for compliance.

6.2.2.2 Mass Market AC Load Control

Table 31 identified three EDCs whose approved Phase III EE&C plan included direct load control of air conditioning equipment. This section of the protocol is titled Mass Market AC Load Control because it is intended to cover both legacy cycling switches and thermostat-based demand response. The key attribute is that the EDC or its CSP has the ability to physically influence the operation of a large number of air conditioning units. Section 5.2 of the 2016 TRM specified that the target sector for DLC was residential and small commercial establishments, and Phase III EE&C plans for several companies include both sectors. Most of the concepts in this section of the protocol could apply to direct load control of other end-uses (water heating, pool pumps, etc.) but the focus is on air conditioning since that is the end-use targeted by EDCs in their Phase III EE&C plans.

The 2016 PA TRM identifies a hierarchy of methods for use in evaluating direct load control and behavior-based Act 129 demand response programs. While an experimental design relying on random assignment to a treatment or control group is really the only defensible method for evaluating behavior-based DR, an RCT approach creates a fundamental issue for mass market AC load control programs. The creation of a control group from program participants necessarily reduces the cost-effectiveness of the programs and harms an EDC's ability to achieve DR goals. Random assignment of accounts to a control group or selecting a sample of accounts to act as the control group for a given event would mean that an EDC and its ratepayers are getting no return (e.g., kW reduction and avoided capacity cost) on its investment of equipment purchase, installation cost, and marketing dollars.

Necessary control group size does not increase linearly with participation, so if an EDC's mass market program is large enough, it may be possible to assign a sufficient number of accounts to a control group to produce precise load impact estimates without seriously harming the expected cost-effectiveness or load impacts. A randomized encouragement design (RED) could be deployed where EDCs only extend the enrollment offer to a subset of accounts, and the remaining accounts are used as a control group. This approach would not have the cost issues associated with installing equipment that does not contribute load reduction, but it would impose an artificial limit on the market size for the program. The SWE recommends that EDCs and their evaluation contractors weight the costs and benefits of using an RCT-like approach that randomizes some participants into test and control groups. The benefits of using an RCT may be considerable for residential and small commercial AC programs. Ultimately, the tradeoff between measurement accuracy and program achievements/cost-effectiveness is an EDC business decision, and the SWE will not require an experimental design for Phase III mass market AC programs because of the program delivery challenges created.

Selection between the two remaining options identified by the TRM (comparison group analysis and within-subjects regression) should be driven primarily by the metering and IT infrastructure an EDC has in place. For EDCs with hourly or sub-hourly meters and the IT capabilities to retrieve the data for analysis for all of the accounts in the target sector, a comparison group analysis is the preferred method because it ensures that event weather conditions are captured in the reference load. Without interval load data for many thousand

non-participating accounts, a comparison group approach via matching is not feasible. Therefore, evaluation contractors would need to rely on an approach that compares interval load measurements of participating accounts on event days to measurements taken on non-event days.

6.2.2.2.1 Comparison Group Analysis

Section 6.2.2.1.1 provided an overview of matching methods and how evaluation contractors may use these techniques to identify a pool of non-participating customers whose event day loads can serve as the counterfactual for DR program participants. Matching or stratified matching is well-suited to the sectors targeted by mass market AC load control programs if there is a large number of relatively homogeneous accounts. This section is intended to build on guidance in Section 6.2.2.1.1 and discuss some of the issues specific to programs that target the air conditioning end-use. The discussion and examples in Section 6.2.2.1.1 focused primarily on the use of binary outcome models for matching, while this section explores an algorithm-based minimum distance approach. This is not to imply that propensity score matching is more appropriate for load curtailment programs and minimum distance algorithms are a better tool for mass market AC. Evaluation contractors may compare the strengths and weaknesses[116] of different methods and select an approach based on their professional judgment.

The primary objective when developing a matched control group for a mass market AC load control program is to find homes that have a similar magnitude and timing of air conditioning usage. If evaluators can identify accounts of non-participants whose loads respond to changes in outdoor temperature similarly on non-event days, it follows that those homes or businesses would behave similarly absent event dispatch. There is a number of different ways to calculate and compare the weather sensitivity of different accounts. Separate comparison could be made by hour and type of day. One approach that is far less data intense is to make an initial pass through the participant and non-participant pool using just monthly billing data and cooling degree days and assign customers to general bins of weather sensitivity. Another option is to stratify customers into summer usage bins using Dahlenius-Hodges or a comparable methodology.

This preliminary stratification allows subsequent steps that are computationally intensive to be performed across smaller subsets of customers. The premise being that customer accounts whose weather sensitivity differs using a coarse metric are not going to be matched anyway, so it is not a prudent use of resources to compare them in detail.

Target hours for Act 129 demand response will generally be hot summer weekday afternoons, so it is important that final matches are based on load similarity on hot non-event weekdays. One approach evaluation contractors may consider is a Euclidian distance

[116] For a useful comparison of pros and cons. Dave Hanna, Kelly Marrin. Control Group Wars - There's More Than One Way to Win the Battle. 2013 International Energy Program Evaluation Conference, Chicago. http://www.iepec.org/conf-docs/conf-by-year/2013-Chicago/044.pdf

calculation.[117] Table 37 presents a hypothetical calculation based on five accounts. Account #1 is a DLC program participant (program=1), and the other four are non-participants (program=0). The data in the top half of the table represent average hourly loads on some set of hot non-event weekdays from noon to 7 pm. The area of interest is a mathematical approach to determine which of the four non-participants is most similar to Account #1. The bottom half of Table 37 shows the distance calculation between Account #1 and the other accounts where distance is the square of the difference between the two values.

Table 37: Euclidian Distance Calculation

Average Hot Weekday Loads (kW)									
Account	**Program**	**HE13**	**HE14**	**HE15**	**HE16**	**HE17**	**HE18**	**HE19**	**Sum (kWh)**
1	1	2.9	3.1	3.8	3.6	4.6	4.4	3.5	25.9
2	0	3.5	3.2	4.0	3.7	5.2	3.7	4.0	27.3
3	0	2.7	4.2	3.5	3.9	5.5	2.8	3.6	26.2
4	0	3.3	4.1	2.0	3.8	5.5	5.1	2.2	26.0
5	0	2.1	2.5	2.2	4.9	4.8	3.7	2.8	23.0
Distance Calculations									
Account	**Program**	**HE13**	**HE14**	**HE15**	**HE16**	**HE17**	**HE18**	**HE19**	**Euclidian Distance**
1	1	0	0	0	0	0	0	0	0
2	0	0.36	0.01	0.04	0.01	0.36	0.49	0.25	1.23
3	0	0.04	1.21	0.09	0.09	0.81	2.56	0.01	2.19
4	0	0.16	1	3.24	0.04	0.81	0.49	1.69	2.73
5	0	0.64	0.36	2.56	1.69	0.04	0.49	0.49	2.50

The rightmost column in the bottom half of Table 37 is the Euclidian distance, which is equal to the square root of the sum of the individual distances. Using this approach, Account #2 would be selected as the best match because it has the smallest distance value. It is worth noting that Account #2 is not the closest match in terms of average energy (kWh) used across the seven-hour period of interest. Accounts #3 and #4 are much closer in terms of total volume but differ in the distribution across hours. This is key for mass market AC load control programs because the occupancy patterns and/or thermostat programming of cooling loads are important drivers of observed load reductions over the course of an event.

In this simplified example a single account was selected as the best match for the participating customer. With large pools of non-participants available, evaluation contractors may choose to select multiple matches for each participating account to increase the stability of the reference load. With any matching approach, it is important to test the accuracy of the approach by excluding a few non-event days from the matching exercise and then comparing the participant and non-participant loads on these days to ensure they

[117] Evaluators may also consider employing Mahalanobis distance matching, which accounts for the covariance between the matching variables and can result in superior matches in comparison to simple Euclidean distance matching.

are well-aligned. Another best practice is to check the balance of covariates across the participant and comparison groups. The two groups should be similar across both energy consumption metrics and descriptive statistics like ZIP code.

The 2016 TRM states that "*difference-in-differences estimators should be used in the analysis to control for any remaining non-event day differences after matching.*"[118] One approach evaluation contractors may consider is a lagged dependent variable (LDV) model that addresses customer uniqueness by including consumption data from non-event weekdays as an independent variable. This could take the form of past days, future days, or some combination of the two. Another technique is to use the customer's load before the hour of control. This will improve the regression results and provide some scaling between customers who are using air conditioning on event days and those that are not.

One of the challenges with aggregate impacts (MW) from mass market AC load control programs is that the participant population is constantly changing. New customers enroll in the program and others exit the program for various reasons. Calculating the total number of active homes or businesses at the time of each event is generally a straightforward calculation, but including every account in the load impact calculations can create challenges. With matching this gets especially tricky because the comparison group homes should ideally come in and out of the data set with their matches. As long as the program population is relatively steady and there are no major directional changes (e.g., multifamily accounts were allowed in after being excluded previously) evaluation contractors may simplify matters by isolating the accounts that were active for the entire summer and using them to develop the comparison group and estimate per-premise load impacts. This average measured value per analyzed participant for each event hour can then be multiplied by the actual number of active accounts to estimate aggregate impacts.

Comparison group analysis of mass market AC load control programs that rely on the installation of new smart thermostats is challenging because the devices can create two types of peak demand impacts.

1) Everyday coincident demand reduction from the efficient operation of the smart thermostat compared to the manual or programmable thermostat installed previously. These impacts do not contribute to DR compliance targets.
2) Event day reductions. The reference load for these impacts estimates should approximate smart thermostat control of the home on a non-event day.

The challenge with implementing a comparison group approach is that most potential comparison group homes will not have smart thermostats, so the event savings estimate could potentially include savings type #1. While this complication could potentially overstate DR savings, evaluation contractors do not need to attempt any adjustments to isolate the event savings when a comparison group approach is utilized for smart thermostats. This recommendation is based on the fact that the Interim Measure Protocol for smart

[118] 2016 Technical Reference Manual. State of Pennsylvania Act 129 Energy Efficiency and Conservation Program & Act 213 Alternative Energy Portfolio Standards. Docket Number M-2015-2469311. Page 524

thermostats[119] does not allocate any energy efficiency peak demand savings to the measure. If evaluation contractors were to net out the IMP demand savings from the estimated DR impacts, the calculation would be a subtraction of zero.

6.2.2.2.2 Within-Subjects Regression

EDCs that offer mass market AC load control programs without AMI for the majority of accounts in the target sector will need to gather interval load data for analysis in one of two ways.

1) Select a representative sample of participating homes or businesses and install interval meters on the home or end-use data loggers on the air conditioning unit(s).
2) Rely on the ability of the load control equipment itself to capture run time of the air conditioning unit.

If EDC evaluation contractors use a metering sample to assess load impacts, the sample should be designed to produce measurements that are accurate to within ± 15% relative precision at the 85% confidence level for each DR event. Ideally, evaluation contractors should leverage EDC load research samples to assess the variability of loads across expected event hours and calculate the coefficient of variation (standard deviation/mean) using interval data. Absent a sample of interval data for the target sector, evaluators may calculate the C_v using billed consumption from summer months and add a cushion to account for the higher level of variability expected in interval data. Evaluation contractors may also consider a stratified design with Neyman allocation to allocate a higher number of sample points to highly variable strata.

If the data loggers deployed measure current, field staff must gather the relevant voltage and power factor measurement to convert amperage to power. Similar data, or proxy variables, are needed if an EDC and its evaluation contractor choose to leverage the data recording capabilities of the load control equipment. This approach is becoming increasingly viable for thermostat-based load control options because the smart thermostat device records when it is calling for heating and cooling and can transmit this information back to the EDC or CSP because it is connected to the home's Wi-Fi.

There are several challenges associated with using run time data from thermostats to calculate DR load impacts that should be addressed in the EM&V plan for the program.

- **Vendor release of data** – Thermostat vendors need to provide customer-level interval run time data to the evaluation contractor for this approach to work, and some vendors are reluctant to share this information. Aggregated run time data from vendors are not adequate for claiming Act 129 gross verified peak demand reductions because the precision of the estimate cannot be calculated.
- **Cooling system specifications** – In a Bring Your Own Thermostat program, no EDC technician visits the home, so gathering actual equipment information needed to convert run time to electric demand can be problematic. Customer self-reports or TRM defaults may be used for capacity and nameplate efficiency.

[119] Approved May 4, 2016

- **Circulating fan savings** – Most homeowners use the auto fan setting on their thermostats, so the circulating fan inside the home's air handler runs only when the system is calling for heating or cooling. Since the smart thermostat reduces demand during events by lowering the amount of time the cooling system operates, it follows that there will be a small reduction in fan power in addition to the savings at the condensing unit. An approach that relies on converting nameplate tonnage and SEER to kW will generally capture this savings because the SEER value reflects fan power. If an assumed voltage and condensing unit amperage calculation is used, evaluation contractors may need to include an assumed fan horsepower and perform a supplemental calculation to capture the fan savings during events.
- **Zero-inflated data** – Depending on the data interval, there may be many records where the load is zero or close to zero. If there are enough such intervals, the data will not be normally distributed and will appear as if the dependent variable is censored at zero or close to zero. In this case, the evaluator should employ a Tobit model to account for the non-normal distribution of the electricity use and to obtain unbiased estimates. It may be possible to aggregate the data to the hourly level to avoid this problem because most HVAC systems will run at least a few minutes in most hours of interest (weekday afternoons).

Equation 21 presented a simple weather-dependent regression model specification that could be used to estimate the reference load for a mass market AC load control program. There are a number of different independent variables and functional forms from which evaluation contractors may select to estimate what the load absent DR would have been. Typically, these models will be pooled where a large number of participants are modeled as a panel to estimate the average impact. Individual participants can be modeled as fixed effects or random effects.

Because DR events will typically be called on the hottest days of the year, evaluators may encounter a situation where there are no non-event days where the weather is as extreme as it is on the event days. Coefficients on the weather variables developed in moderate temperature ranges may not completely capture the relationship between load and temperature at extreme temperatures. One key difference between load curtailment programs and mass market AC load control programs is that, in the latter, participants are typically not notified of upcoming events. While same-day adjustments are prohibited for load curtailment programs, they can be a useful tool for mass market AC load control programs to supplement the weather term(s) and calibrate the reference load to extreme weather conditions. If a same-day adjustment term is used in the regression model specification, evaluation contractors should be mindful of vendor pre-cooling algorithms and make sure the adjustment term is developed from data prior to the beginning of any pre-cooling.

Section 6.2.2.1.2 discussed the differences between using regression for ex post analysis and ex ante forecasting of impacts. Once multiple load control events have been called at different temperatures, evaluation contractors should consider developing a time-temperature matrix (TTM) of expected load impacts by hour and outdoor air temperature for planning purposes. To produce a TTM when a comparison group approach is used,

evaluators may fit a second stage model comparing estimated load impacts to weather conditions.

6.2.2.3 Behavioral Demand Response

Table 31 indicated that five of the seven EDCs intend to offer behavioral demand response (BDR) programs in Phase III. These programs are similar to the Home Energy Report (HER) programs in that they seek to modify customer energy consumption through a combination of energy-saving recommendations and information about usage and how it compares to other homes. Unlike the HER programs, behavioral demand response programs will target reductions during specific hours (DR event days). These programs will also rely more on electronic media and communications than HER programs because there simply is not sufficient time to print and mail a report from the time an event day is determined to the start of the event. Instead, conservation messaging will rely on email, text messages, and other electronic communications to engage participants.

Unlike mass market AC load control programs, the EDC will not have any physical control over the loads within the home. Unlike load curtailment programs, participants will not be compensated for the reductions they produce (other than any bill savings). The expected savings from behavioral DR in Phase III EE&C plans range from 50 to 80 Watts per home. While the savings are modest, the program delivery costs are limited because there are no equipment or incentive expenses. In order to produce measurements of demand reductions that are statistically significant, behavioral DR programs will need to use a combination of large sample sizes and sound experimental design. Ideally, hourly or sub-hourly metering would be available for all homes in the program. If an EDC enters the PY9 demand response season with incomplete interval meter deployment for the BDR program population, evaluation contractors may propose an approach to estimate the BDR program impacts using only those homes with interval meters installed. The proposal should include a validation exercise to assess whether there are any systematic differences between the AMI and non-AMI population that could potentially bias impact estimates.

6.2.2.3.1 Experimental Design

BDR programs should rely on an experimental design where randomization is used to create equivalent treatment and control groups. The treatment group receives the program messaging encouraging participants to shed load during event hours, and the control group's loads serve as the reference load. An evaluation of PG&E's BDR program during the summer of 2015 found that "non-random differences between the treatment and control group become apparent on both event and non-event days."[120] The presence of savings on non-event days indicates that an experimental design where the two randomized groups alternate treatment and control status from event to event would likely understate impacts, so it is not recommended for Act 129 BDR programs.

One key experimental design issue EDCs, vendors, and evaluation contractors need to be particularly cognizant of is the overlap between the BDR population and the HER program

[120] http://www.calmac.org/publications/Behavioral_Demand_Response_Study_Final_Report_CALMAC.pdf, p. 20.

population. A design where the HER treatment group acts as the BDR treatment group and the HER control group serves as the BDR control group is flawed because estimates would capture both the BDR impacts and the coincident demand reductions achieved by the HER program.[121] Two acceptable solutions would be as follows:

1) Create two separate cohorts within the BDR program. One treatment/control cell would be created within the HER program treatment group, and a second randomization would occur among homes that do not receive HERs (either from the HER control group or homes that are not in the HER program at all).
2) Randomly select a BDR treatment and control group from the eligible population without consideration of HER status. As part of the randomization validation, run an equivalence test to confirm that the two groups contain similar proportions of HER treatment group homes.

Regardless of the selected method, evaluation contractors should be mindful of introducing potential complications for the evaluation of the HER program when designing the BDR deployment. Because of the small expected impact and noisiness of hourly load data, precision will be a challenge for BDR programs. EDCs should design group sizes in a way that the expected margin of error associated with the per-home impact estimate at the 95% confidence level for a single event is no larger than the expected reduction (e.g., statistical significance).

6.2.2.3.2 Model Specification

Even with large sample sizes and proper randomization, subtle differences between the treatment and control groups may exist that could bias BDR results from a simple difference in means calculation. EDC evaluation contractors should consider a regression framework that addresses the uniqueness of customers, either through participant-level fixed effects or lagged demand terms. The LFER, LDV, and LS model specifications discussed in the behavioral protocol can, with a few modifications, be used to estimate BDR impacts. Evaluation contractors should include the model specification that will be used to estimate gross verified demand savings in the EM&V plan for the program.

Based on findings from the PG&E BDR evaluation as well as a recent Hydro Ottawa BDR evaluation[122] stating that BDR savings appear on non-event days as well as event days, model specifications should not include lag terms from the intervention period because the inclusion of these terms could bias impact estimates from event days downward.

6.2.3 Uncertainty

Estimating demand response impacts is an inherently counterfactual exercise. The energy consumption can be measured, but any reduction in load needs to be estimated by establishing a baseline or reference load. On event days, what the load absent dispatch

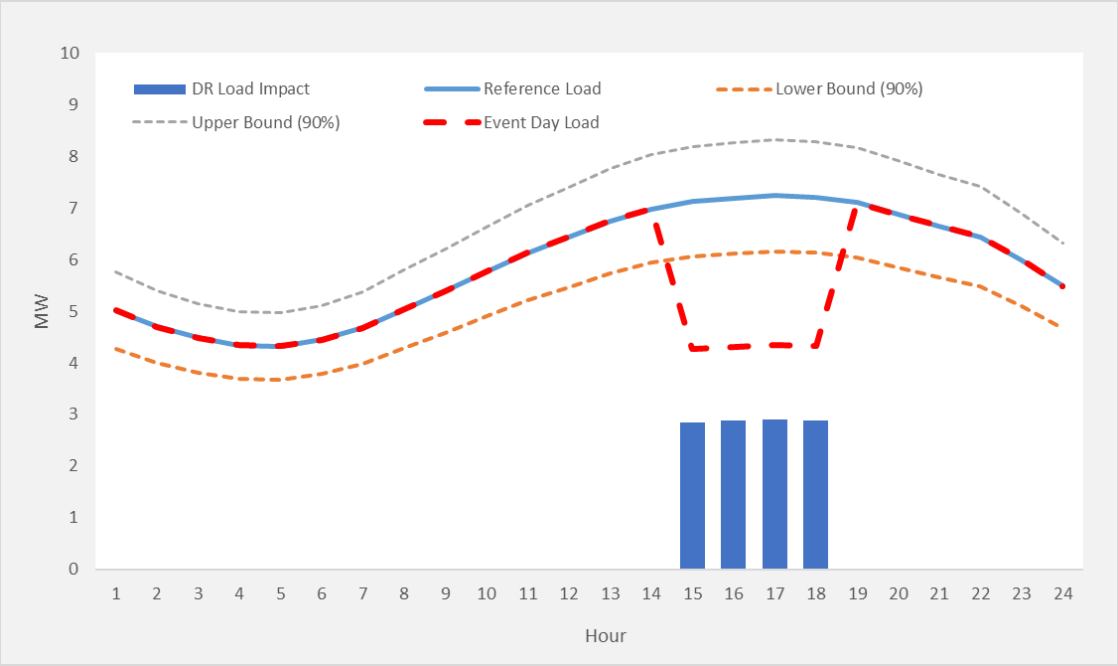
[121] Assuming a flat load shape and a per-home savings of 200 kWh annually, the HER peak demand reduction would equal 22 Watts or ~30% of the expected BDR savings.
[122] http://www2.opower.com/l/17572/2016-02-17/3md2yb/17572/122339/Hydro_Ottawa_Behavioral_Demand_Response_Evaluation___Final.pdf

would have been cannot be measured and must be estimated (e.g., the reference load). Likewise, on non-event days, what the load would have been if a DR event had been called cannot be measured and must be estimated. Like any estimate, there is some level of uncertainty in the reference load estimate that is a function of the amount of variability in loading patterns as well as any sampling conducted by the evaluation contractor. This uncertainty band is typically referred to as the margin of error—or precision—of the estimate, and can be expressed on either an absolute (kW) or relative (%) basis.

It is important for evaluation contractors to calculate and report the margin of error associated with demand response load impacts because it provides stakeholders with a quantitative sense of how precise the reported load impact estimates are. The SWE will also take the margin of error into consideration in any recommendations to the PUC about compliance or penalties for failure to achieve statutory DR goals in its Phase III reports—although compliance determination is ultimately at the sole discretion of the Commission. This section provides evaluation contractors with guidance on how to calculate and report uncertainty and is organized by estimation method. For Act 129 demand response programs, uncertainty should be expressed at the 90% confidence level, and all precision values in the section are presented at 90% unless otherwise noted.

Demand response load reduction estimates are equal to the difference between an estimated reference load and the observed load, and there is typically no uncertainty in the observed load because it is metered. This means that the margin of error of the load reduction estimate is equal to the margin of error of the reference load when expressed in absolute terms. However, there is a key difference between the margin of error of the reference load and the margin of error of the load reduction estimate when examined on a relative basis. Consider the example shown in Figure 15. The margin of error of the reference load when it is equal to 7 MW is approximately ± 1 MW (or ± 15%). The facility in the hypothetical example below reduces load by approximately 3 MW during each of the four event hours. The margin of error of the load reduction estimate is also equal to ± 1 MW, but when the margin of error is expressed on a percent basis, the result is 1/3 or a margin of error of ± 33.3%. The precision of the load reduction estimate is the parameter of interest for Act 129 DR programs.

Figure 15: Demand Response Margin of Error Example



6.2.3.1 Regression Analysis

Regression-based impact evaluation approaches can be complex to implement, but the uncertainty calculation is fairly straightforward once the model specification is selected and estimated. The standard error of the impact coefficient is the key parameter of interest whether a within-subjects or comparison group design is selected. The impact coefficient represents the least squares or maximum likelihood estimate of the relationship between the independent variable and dependent variable, and the standard error is a measure of how precisely the model estimates that relationship.

Equation 23 shows the simplest form of a site-specific regression model. Metered load in hour h is the dependent variable. The *Event* term is an indicator variable equal to 1 if hour h is a DR event hour and zero otherwise. β_0 is the model intercept and represents the average load absent a DR event, and β_1 is the regression coefficient for the event indicator and represents the estimated load impact during a DR event. The sum of β_0 and β_1 thus equals the estimated facility load during a DR event.

Equation 23: Individual Customer Regression Model

$$kW_h = \beta_0 + \beta_1 * Event_h$$

In this simplified example, there is a single event day, and the model is estimated using data from hour 17 on summer weekdays (n=84 days). The β_1 coefficient indicates that the model estimates a load reduction of 334.2 kW during the event.

Table 38: Sample Regression Output

Source	SS	df	MS
Model	110,342	1	110,342
Residual	144,880	82	1,767
Total	255,222	83	3,075

Number of obs	84
F(1, 82)	62.45
Prob > F	0.000
R-squared	0.4323
Adj R-squared	0.4254
Root MSE	42.034

kwh	Coef.	Std. Err.	t	P>|t|	Lower Bound 90% CI	Upper Bound 90% CI
Event (β_1)	-334.2	42.29	-7.9	0.000	-404.5	-263.8
Intercept (β_0)	535.8	4.61	116.13	0.000	528.1	543.5

The standard error value for the β_1 coefficient (42.29) is then multiplied by the t-statistic for the desired level of confidence to calculate the margin of error around the impact estimate. In this example, the margin of error is ± 70 kW, or 21% of the impact estimate at the 90% confidence level.

Although standard errors are always in the same units as the response variable, when interaction terms are used in more complex specifications, an extra step is created for evaluation contractors. For weather-dependent regression models, like a mass market AC load control program, evaluation contractors will often choose to interact an event indicator variable with a weather term like outdoor air temperature. With this type of specification, the impact coefficient and its standard error will be expressed as a function of temperature and need to be multiplied by the observed event conditions to calculate the load impacts and the associated margin of error.

When calculating and reporting the precision of demand response impacts via regression, it is important to be mindful of what the indicator variable represents and therefore what the margin of error surrounds. Depending on the specification of the model and the observations included, impact coefficients and their standard errors could be estimating any of the following:

- An individual DR event hour
- Individual event indicator
- Program Year event indicator
- Phase III event indicator

For a site or program where DR performance is relatively consistent, standard errors should improve as more event data are available to estimate a regression coefficient. This means that the margin of error for an *average* event impact estimate should be tighter than the margin of error around a *single* event. Because the PUC established performance goals for both the Phase III event average and each individual event, evaluation contractors will need to estimate several model specifications to obtain the necessary standard errors for reporting. Table 39 illustrates what uncertainty reporting might look like in Program Year 10 once impacts are being aggregated across events and previous program years. Section 6.2.3.3 describes the aggregation of errors across multiple participating sites.

Table 39: Sample Uncertainty Reporting Table for PY10

DR Event	Load Reduction Estimate (System Level MW)	Relative Precision (90% Confidence)
July 12, 2018	51	18%
July 21, 2018	54	17%
August 2, 2018	50	21%
August 3, 2018	57	19%
PY10 Average Performance	**53**	**11%**
Phase III Average Performance	**52**	**9%**

Special consideration is required for regression models that include multiple program participants. Section 6.2.2 discussed how pooled[123] regression models can be an efficient way to estimate the average or aggregate DR impact across a large number of DR program participants. Pooled models violate one of the key assumptions of ordinary least squares regression that model errors are independent. Typically, in a pooled model we find that errors are clustered within observations from a given participant, or that they have heteroscedasticity. Consider a small home in an AC load control program. If the home is smaller than the typical home with concordant lower average demand, the regression model will consistently overestimate the load in the home, and the errors will be clustered around the average difference in load.

Using default standard errors with a pooled regression model will generally overstate the precision of the estimate. Instead, statistical inferences about the precision of the load reduction estimate should be based on cluster-robust standard errors, which are based on the variability observed between participants. Most statistical packages offer the option to produce robust standard errors. Evaluation contractors just need to identify the cluster variable, which will typically be the EDC account number or some other unique identifier for the DR participant.

6.2.3.2 Day Matching

Day matching techniques, or CBLs, do not produce standard errors, so an alternative approach is needed to estimate the uncertainty associated with the reference load and load impact estimates. Day matching techniques are the least robust calculation method and will typically only be used for load curtailment participants with non-weather-dependent loads who have an approved alternative CBL registration with PJM. PJM uses relative root mean square error (RRMSE) as a metric of uncertainty, so these accounts will have a convenient proxy statistic that can be utilized for Act 129 uncertainty calculations. PJM Manual 11[124]

[123] Also referred to as cross-sectional or panel models
[124] http://www.pjm.com/~/media/documents/manuals/m11.ashx, p. 130.

requires 60 days of contiguous non-event load data and specifies the RRMSE calculation as follows:

- To perform the RRMSE calculation, daily CBL calculations are first performed for the CBL method using hours ending 14 through hours ending 19 unless otherwise approved by PJM as the simulated event hours for each of the 60 non-event days according to the CBL method rules.
- Actual Hourly errors are calculated by subtracting the CBL hourly load from the actual hourly load for each of the simulated event hours of the non-event day.
- The Mean Squared Error (MSE) is calculated by summing the squared actual hourly errors and dividing by the number of simulated event hours.
- The Average Actual Hourly Load is the average of the actual hourly load for each of the simulated event hours.
- The Relative Root Mean Squared Error (RRMSE) is calculated by taking the square root of the MSE, then dividing that quantity by the average of the actual load.

The RRMSE statistic represents the percent error associated with the reference load, so an additional step is required to calculate the uncertainty of the load reduction. Table 40 illustrates the precision calculation for a hypothetical customer.

Table 40: Sample Precision Calculation Using RRMSE

Impact Statistic	Value
Reference Load	500 kW
RRMSE	15%
Margin of Error of Reference Load	75 kW
Load Reduction	150 kW
Relative Precision of Load Reduction	50%

EDC evaluation contractors can establish the RRMSE value associated with each account on an annual basis and use the static calculated value in the uncertainty calculations for all event and program year totals for that account. Although the RRMSE statistic does not have an associated confidence level, for Act 129 reporting purposes the percent error can be assumed to represent the 90% confidence level for simple aggregation with other sources.

6.2.3.3 Aggregation of Errors

The RRMSE-based calculations described in Section 6.2.3.2 will result in a separate uncertainty calculation for each DR participant where day matching techniques are used. A series of individual customer regressions described in Section 6.2.3.1 will produce a similar data set. Aggregation of the load reductions across participants is a simple sum of their individual performance estimates. The aggregation of errors should be calculated using the

square root of the sum of squared individual absolute margins of error. Table 41 illustrates the calculation for three hypothetical load curtailment participants.

Table 41: Aggregation of Participant Level Errors

Parameter	Site #1	Site #2	Site #3	Total
Reference Load (kW)	5,000	5,000	3,000	13,000
Load Reduction (kW)	1,000	1,000	1,000	3,000
Absolute Margin of Error (kW)	500	1000	150	1,128
Relative Precision of Load Reduction (+/-)	50%	100%	15%	38%

The margin of error of the aggregation load reduction is equal to:

$$\sqrt{500^2 + 1000^2 + 150^2} = 1{,}128\ kW$$

And the relative precision of the aggregate load reduction is equal to:

$$\frac{1{,}128}{3{,}000} = \pm\ 38\%$$

6.2.3.4 On-Site Generation

When load reduction estimates are based on metered output from behind the meter generation, there is effectively no uncertainty in the load reduction because the output is assumed to displace grid-supplied load 1:1. For this type of load curtailment participant, EDC evaluation contractors should assume a margin of error of ± 0 kW. This assumption holds true only for sites that do not self-generate on non-event (PJM or Act 129) days.

6.2.3.5 Sampling

For some demand response programs, EDC evaluation contractors may choose to analyze a sample of program participants and extrapolate findings to the population. Sampling may be necessary if the installation of end-use metering equipment is needed or if all participants do not have interval meters. Sampling may also be used within a load curtailment program to conserve evaluation resources.

When sampling is used in combination with pooled regression methods, the standard error(s) of the impact coefficients encompass the sampling error, so no additional correction is necessary. This useful feature stems from the fact that the magnitude of the standard error is determined, in part, by the number of observations, or clusters of observations, used in the impact estimate. As sample size decreases, the modeled standard errors will increase and account for the uncertainty introduced by sampling.

When sampling is used in combination with individual customer regressions or day matching techniques, a supplemental correction is needed to account for the sampling error. Calculation of the estimation error within the participant sample would follow the methodology described in Sections 6.2.3.1 and 6.2.3.2. Calculation of the sampling error for

a realization rate or mean-per-unit estimate follows the same calculation steps for demand response as for energy efficiency. The propagation of error when the reported savings of the program population is multiplied by the parameter of interest relies on the sum of squared relative error calculation shown in Equation 24.

Equation 24: Propagation of Error Formula

$$Relative\ Error\ of\ Verified\ Savings = \sqrt{(RP_A)^2 + (RP_B)^2}$$

The RP_A term is the relative error associated with the individual customer estimation method(s) used for the sample. The RP_B term is the relative error introduced by analysis of a sample of participants rather than a census. Sample size and the degree of correlation between ex ante and ex post savings estimates determine the magnitude of the sampling error. Table 42 presents a sample calculation for a hypothetical load curtailment program.

Table 42: Propagation of Error Example

Evaluation Parameter	Value	Relative Error	Absolute Error
Reported kW Savings	50 MW	12.0%	6 MW
Realization Rate	90%	5.0%	4.5%
Verified Savings	45 MW	13.0%	5.85 MW

6.2.4 Cost-Effectiveness

The 2016 TRC Order[125] provided guidelines for calculating the benefits and costs of Phase III demand response programs. This section of the protocol summarizes key technical issues and discusses the practical application for different program types.

6.2.4.1 Benefits

To calculate the benefits from demand response programs, "*EDCs would average the gross verified demand reductions over each hour of performance and apply a line loss adjustment factor to estimate the magnitude of the peak demand reduced. This demand reduction value would be multiplied by either two or three avoided cost-of-capacity values depending on customer sector."*[126] Table 43 shows the appropriate types of capacity to monetize by sector.

[125] Final 2016 TRC Test Order. Docket No. M-2015-2468992. Entered June 22, 2015. http://www.puc.pa.gov/pcdocs/1367195.docx
[126] Ibid., p. 52.

Table 43: Avoided Capacity Types by Sector

Capacity Type	Residential	Small C&I	Large C&I
Generation	✓	✓	✓
Transmission	✓	✓	✓
Distribution	✓	✓	✗

The TRC Order also stated that avoided energy costs could be used as benefits in the TRC Test. These kWh savings during event hours will be available from the impact evaluation, and evaluation contractors may extend the period of examination to surrounding hours to capture the full distribution of energy impacts. The TRC Order also allows EDCs to implement the assumption used by the SWE in the DR Potential Study "*where each kWh reduced during a DR event was offset by an extra kWh used during an off-peak hour. Using this approach, the avoided cost of energy attributable to a DR program would be equal to the kWh impact during event hours multiplied by the difference in the EDC's on-peak and off-peak summer avoided cost of electricity for the program year.*"[127]

For mass market AC programs, the 2016 TRC Order specified a 10-year measure life for load control equipment and directed EDCs to base cost-effectiveness calculations on "*benefits and costs which have occurred, or which are known to be likely to occur throughout the life of the DLC equipment.*"[128] The practical implication of this directive is that, in each year of Phase III, EDCs will amend the inputs of the Phase III TRC test to include more actual data and fewer projections of cost and benefit. Estimates of future demand reductions are an ideal application of the time-temperature matrix exercise described in Section 6.2.2.2.2.

6.2.4.2 Costs

Quantifying incremental cost is challenging for demand response because participants typically forego comfort or production in order to achieve peak demand reductions, and it is difficult to place a dollar value on what is sacrificed. The Act 129 incentive payment[129] itself is a cost to the EDC and a benefit to the participant, but the incentive payment is also generally considered to be a reasonable proxy for the "cost" of these sacrifices to the participant. The Commission adopted a 75% participant cost assumption for Phase III demand response programs. This assumption is based on the premise that DR incentives likely outweigh the costs of participation for the average participant or else they would not enroll in the program. If an EDC pays each participant $50 to participate in a mass market AC load control program for the summer, $37.50 would be used as cost in the denominator of the TRC Test. When EDCs engage a CSP to aggregate DR customers, visibility into incentive payments is lost. The 2016 TRC Order allowed EDCs to use 75% of the payment to CSPs as a simplifying assumption. If the CSP payment includes both incentives and

[127] Ibid., p. 53.
[128] Ibid., p. 59.
[129] Incentive payments from PJM are not included in the TRC Test for Act 129.

purchase and installation of DLC equipment, EDCs should attempt to separate the cost categories and use the full equipment and installation cost in the TRC Test.

Behavioral DR is an interesting offering from a TRC perspective because there are no customer incentives. While there are likely costs for participants, evaluation contractors do not have a way to quantify them, and therefore the participant costs are not included in the TRC. The TRC costs for a BDR program are simply the program administrator costs of the EDC and fees paid to the program CSP. The inclusion of participant costs for programs paying incentives but exclusion of them for behavioral DR makes comparison of cost-effectiveness between the two program types difficult and potentially disadvantages programs that pay incentives.

Other notable guidelines on estimating DR costs in the 2016 TRC Order are listed below.

- The cost of DLC equipment purchased outside of an approved Phase III plan should not be included as cost
- DR resources that clear as wholesale resources in PJM markets should use the actual financial compensation received instead of the avoided cost of generation capacity and energy calculation methods described in the 2016 TRC Order

6.2.5 Process Evaluation

The SWE recommends that the EDCs conduct process evaluations in order to support continuous program improvement in their DR programs. The SWE recommends following the process evaluation guidance in Section 3.5 to identify opportunities for improvement and successes that can be built upon.

6.2.6 Reporting

Act 129 DR events can only be called from June through September, which are the first four months of the Act 129 program year (June 1 to May 31). The front-loaded nature of the DR season within the program allows for earlier reporting of gross verified impacts from demand response than from energy efficiency programs. Table 44 lists the key activities and deliverables for demand response programs with associated dates for Program Year 9. The cycle would repeat for PY10 to PY12.

Table 44: PY9 DR Reporting Schedule

Milestone	Estimated Date
EDC demand response events occur	June – September 2017
EDC evaluation contractors collect load data and estimate load impacts	October – December 2017
EDC reports gross verified demand reductions PY9 semiannual report to the PUC	January 15, 2018
SWE Team issues PY9 DR data request	January 15, 2018
DR data request response provided to the SWE	March 1, 2018
SWE Team verifies load impact estimates and performs independent estimates as needed	Spring 2018
SWE Team submits PY9 update report to the PUC summarizing EDC gross verified demand reductions from PY9	August 15, 2018
EDC summarizes the final PY9 verified savings in annual report to the PUC. Update savings claims if necessary based on SWE audit findings	November 15, 2018
SWE memorializes PY9 verified peak demand reductions in SWE annual reporting to the PUC	February 28, 2019

EDC reporting templates include specific tables and figures, but the key outcomes are summarized in Table 45. EDC evaluation contractors will provide a load reduction estimate (kW or MW) for each DR program for each DR event hour. The sum of program impacts yields the performance of the DR portfolio for the event, and the average across events returns the average performance for the program year.

Table 45: Sample Demand Response Reporting Template

Event Date	Start Hour	End Hour	Small CI Load Curtailment	Large CI Load Curtailment	Residential DLC	BDR	Average MW Impact
July 12	15	18					
July 25	15	18					
August 3	14	17					
Average PYX DR Event Performance							
Average Phase III DR Event Performance							

Evaluation reports should also include the following elements:

- A summary of the evaluation methodology used to estimate load impacts

- A comparison of observed impacts with planning estimates or participant commitments
- A discussion of any challenges or recommendations for program improvement

Section 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 in order 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 and confidence.

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 a Performance Evaluation Group of the EDCs, their evaluation contractors, 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 in order to mitigate risks to the EDCs. The common goal requires that kWh/yr and kW savings be clearly defined, auditable, and provide a sound engineering basis for estimating energy savings.

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 also could 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 will include 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 also used to describe the customer billing and payment streams associated with customer accounts. This term is used to describe both consumption and 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 ($Cv = \sigma/\mu$).

CONFIDENCE: An indication of how close a value is to the true value of the quantity in question. A confidence interval (CI) 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 Electric Distribution Company (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.

Cv: See Coefficient of Variation.

DEEMED SAVINGS: Technical Reference Manuals (TRM) 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 (equals 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. Demand response programs may include contractually obligated or voluntary curtailment, direct load control, 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, CFLs, LEDs, linear fluorescents, air-source heat pump (ASHP), 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): Energy efficiency and conservation 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 heating, ventilation, and air conditioning (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 program Implementation Conservation Service Providers (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 nonparticipant 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 kWh/yr, kW, and therms) and/or behavior attributed to energy efficiency and demand response programs.

IMPLEMENTATION CONSERVATION SERVICE PROVIDERS (ICSP): Contractor 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 practice 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 an 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 nonparticipant 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.

NONPARTICIPANT: 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 "nonparticipant" 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.

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 2016-2021.

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 EVALUATION GROUP (PEG): Created by the PUC to, among other things, provide guidance to the SWE in clarifying energy savings measurement protocols and plans by recommending improvements to the existing TRM and other aspects of the EE&C 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) to 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.

SIMPLE ENGINEERING MODEL (SEM): A category of statistical analysis models that incorporate the engineering estimate of savings as a dependent variable.

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 nonparticipant 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 III EE&C programs implemented by the seven EDCs in Pennsylvania subject to the requirements of Act 129.

STATEWIDE EVALUATION TEAM (SWE TEAM): The team, led by NMR Group Inc., that is conducting the evaluations of the Phase III Act 129 programs. Team members are NMR Group Inc., EcoMetric Consulting LLC, Demand Side Analytics LLC, Optimal Energy, and Abraxas Energy Consulting.

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., net-to-gross ratio values), source documentation, specified assumptions, and other relevant material to support the calculation of measure and program savings—and the application of such values and algorithms in appropriate applications.

TECHNICAL WORKING GROUP (TWG): Chaired by PUC staff and comprised of representatives from the EDCs, the SWE, and other interested parties to encourage discussions of the technical issues related to the EM&V of savings programs to be implemented pursuant to Act 129.

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

Appliance retirement programs (ARP) typically offer some mix of incentives and free pickup for the removal of old-but-operable refrigerators, freezers, or room air-conditioners. These programs are designed to encourage the consumer to:

- 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 Uniform Methods project 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.[130] It is important to note that appliance replacement (with early retirement) programs are extensions of appliance retirement programs. 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.

B.1 General Free Ridership Approach

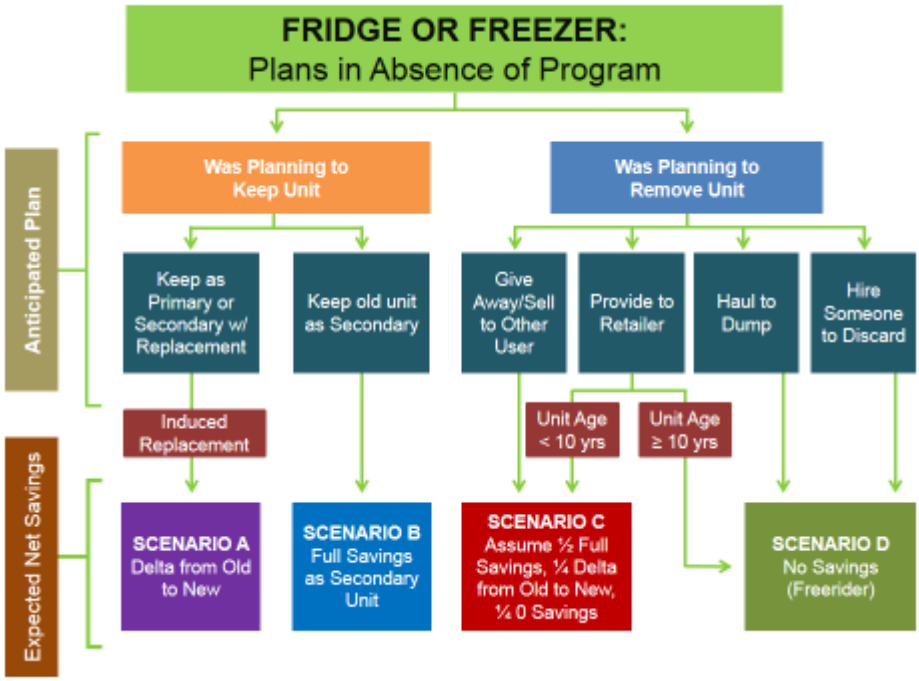
The nature of the appliance retirement program requires a unique approach to estimating free ridership, and ultimately, net savings. Free ridership is based on the participants anticipated plans had the program not been available – a free rider is classified as one who would have removed the unit from service irrespective of the program. Net savings for the appliance retirement program is therefore based on the participants' anticipated continued operation of the appliance either as primary or secondary unit, within their home or transferred to another home (either directly or indirectly).

The general approach to estimating net savings for the appliance retirement program is a several-step process to segment the participants into different groups, each with unique savings attributable to them. Participants should first be classified as either "keepers" or

[130] See The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures, Chapter 7: Refrigerator Recycling Evaluation Protocols, National Renewable Energy Laboratory, March 2013 (Download available at http://www1.eere.energy.gov/wip/pdfs/53827-7.pdf).

"removers." The "keepers" segment, defined as those who had been planning to keep the unit, should be further segmented into groups based on whether they replaced the unit. The "removers" segment, defined as those who had been planning on removing the unit, should be further segmented into groups based on whether the unit would have continued operating or would have been permanently removed from service. Each respondent is then assigned a net savings value, and overall program net savings is calculated in aggregate across the sample population. A simple flow chart, included below in Figure 16, shows how the net savings are derived for the appliance retirement program. A more detailed discussion follows below.

Figure 16: Diagram to Determine Appliance Retirement Net Savings



B.2 CLASSIFYING THE PARTICIPANT AS "KEEPER OR REMOVER"

The first step is to classify each participant as a keeper or remover. This first classification is assessed through a series of questions used to determine what the participant likely would have done if the appliance had not been removed by the program. The following example shows the basic approach:[131]

1. Were you using the recycled unit as your primary [appliance], or had it been a secondary or spare?
   a. Primary
   b. Secondary

[131] Note that these questions are provided as examples of questions to derive the information needed to classify participants into the various scenarios. EDCs can adapt these questions as long as they can provide the same information.

2. If the appliance pickup program was not available, would you have still removed the [appliance], or would you have kept it?
   a. Removed it
   b. Kept it

B.3 CLASSIFYING THE STATUS OF THE "KEEPER"

The "keepers" segment, as discussed previously, would consist of those who answered question 2 above as "b. kept it." The keepers segment is not qualified as free riders, assuming that, in absence of the program, their [appliances] would have continued operating normally. These respondents should be further segmented into groups based on whether they replaced their unit – this helps define the deemed savings values to be assigned to them. The following question is an example of what to ask:

3. Did you replace the removed [appliance] with a different unit?
   a. Yes, replaced the unit (Scenario A, below)
   b. No, did not replace the unit (Scenario B)

The "keeper" respondents who indicate that the [appliance] was not replaced with a different unit (are assigned the full savings (scenario B below). The "keeper" respondents who indicate that the [appliance] was replaced with a different unit and that the replacement is determined to have been induced by the program are assigned the replacement TRM-based deemed savings values. This is typically a small percentage of participants, however, as the incentive usually covers only a very small percentage of the incremental cost of purchasing a new unit. The following set of questions helps to further classify the net replacement savings based on the replacement type of unit.

4. Did you replace the removed [appliance] with another one?
   a. Yes
   b. No

5. Is the replacement [appliance] an ENERGY STAR or high efficiency model?
   a. Yes
   b. No

6. Was this replacement [appliance] brand new or used?
   a. Brand New
   b. Used

As previously mentioned, the proportion we expect to have been induced is very small, so it is critical that the respondents answer is repeated to them and asked again; clarifying questions include the following:

7. Would you have purchased your replacement [appliance] if the recycling program had not been available?

8. I would like to confirm your answer, are you saying that you chose to purchase a new appliance because of the appliance recycling program, or are you saying that you would have purchased the new [appliance] regardless of the program?

If the respondent confirms that they would have purchased the new [appliance] regardless of the program, then by definition they cannot have been induced to purchase by the program and are not classified as scenario A (induced replacement). At this point we still need to determine what they would have done with the old unit. You can then ask if they would have kept it (classified as scenario B) or removed it (continue on to next section regarding "remover"). If the respondent confirms that they would not have purchased the replacement unit without the program then they are considered an induced replacement and get the appropriate replacement savings value from the TRM.

B.4 CLASSIFYING THE STATUS OF THE "REMOVER"

The "remover" segment, as discussed previously, would consist of those who answered question 2 above as "a. remove it." The remover segment is potentially qualified as free riders, assuming that, in absence of the program, their appliance would have been removed from service. These respondents should be further segmented into groups based on whether the unit would have continued operating or would have been permanently removed from service – this helps define the deemed savings values to be assigned to them. The following questions are an example of what to ask:

9. If the appliance pickup program was not available, which one of the following alternatives would you have most likely done with your [appliance] when you were ready to dispose of it? Would you have:

    a. Sold it
    b. Given it away for free
    c. Had it removed by the dealer you got your replacement [**appliance**] from
    d. Took it to a dump or recycling center
    e. Hired someone else to haul it away

Ask the following question if the answer to question 9 above is "a – Sold it":

10. You said you would have most likely sold your [appliance]. Would you have sold the [appliance] to an appliance dealer, or to a private party (like a friend, relative or by running an ad)?

    a. Dealer
    b. Private party (friend, relative, or by running ad)

If the anticipated plan was to sell the unit to a dealer, then the age of the unit needs to be discovered. Ask the following question if the answer to question 9 above is "a. Dealer":

11. You said you would have most likely sold your [appliance] to a Dealer. Was your [appliance] less than 10 years old?

    a. Yes, less than 10 years old
    b. No, at least 10 years old

We can assume that operable units less than 10 years old (answer "a. Yes, less than 10 years old" to question 11) are likely to be resold on the open market (qualified as scenario C), whereas units at least 10 years old (answer "b. No, at least 10 years old" to question 11)

are likely to be removed from service (scenario D) since there is little probability of them having any retail value greater than the cost of attempting to sell the unit.[132]

Ask the following question if the answer to question 9 above is "b – Given it away for free":

12. You said you would have most likely given away your [appliance]. Would you have given it to a private party (like a friend, relative or by running an ad), or to a charitable organization?

    a. Private party (friend, relative or by running an ad)
    b. Charitable organization

Ask the following question if the answer to question 9 above is "d – Took it to a dump or recycling center":

13. You said you would have most likely taken away the [appliance] yourself. Would you have taken it to a dump, or to a recycling center?

    a. Dump
    b. Recycling Center

If the respondent was planning on transferring the unit by selling (if unit is less than 10 years old) or giving it away (answers a or b for question 9), then we can assume the unit would likely continue operating and therefore the respondents are not classified as free riders. The savings attributable to these participants are the most difficult to estimate, because this scenario (Scenario C) is attempting to estimate what the prospective buyer of the used appliance did in absence of finding the program-recycled unit in the marketplace (i.e., the program took the unit off the grid, so the prospective purchaser faced, in theory, a smaller market of used refrigerators). The UMP uses, and, in absence of primary data collection, this guideline recommends, a composite value for this scenario, assuming one-half of the respondents would receive full savings (assuming unit would have served as secondary unit for a different household), one-quarter of the respondents receive the delta between a new and old unit (non-ENERGY STAR), and the remaining one-quarter of the respondents receive zero savings (assuming different household was able to find alternative similar old unit).

If the respondent was planning on removing the unit from service, either through recycling it, hauling it to the dump, or hiring someone to remove it (answers c, d, e to question 9), or if they planned on giving it to a retailer but the unit was at least 10 years old, then they are classified as full free riders and not allocated any savings (scenario D, net savings = 0).[133] One final consideration with respect to the free rider scenario is the availability of disposal options in the service area in question. Evaluators may want to include viability/logistics of alternative options (whether there is even possibility of this service in participant's area) in

[132] The 10-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

[133] Scenario D assumes that the retailers that picked up the unit would have discarded the unit, rather than selling it on the secondary market.

advance of fielding the survey. If it is discovered that no such option exists, then additional options need to account for alternative possibilities in the survey.

B.5 Estimating Net Savings

Net savings should be assigned individually to each respondent based on the responses provided to the questions outlined above. The net savings should be averaged across all respondents to calculate program-level net savings. Table 46 demonstrates the proportion of a sample population that are classified into each of the potential seven categories and the resulting weighted net savings.

Table 46: Net Savings Example for a Sample Population*

Primary Classification	Secondary Classification	Replacement TRM value	Population (%)	UEC (kWh) w/out Program	UEC (kWh) w/ Program	kWh Savings
Would have kept unit	Scenario A: Induced Replacement	Non-ES unit	3%	1,026	520	506
	Scenario A: Induced Replacement	ES unit	2%	1,026	404	622
	Scenario B: Secondary unit w/out replacement	No replacement	25%	1,026	0	1,026
Would have removed unit	Scenario D: Removed from service	No replacement	20%	0	0	0
	Scenario C: Transferred	No replacement or unit age >= 10 years	12.5%	0	0	0
		Non-ES unit, unit age < 10 years	12.5%	1,026	520	506
		No replacement	25. 0%	1,026	0	1,026
Net Savings (kWh)						**604**

* 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 2016 PA TRM to determine the UEC of retired units.

B.6 Data Sources

A random sample survey of program participants should be the primary source of data collected for estimating net-to-gross 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 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 net-to-gross ratio (NTG) for Phase III of Act 129 is to be treated in the same way as for Phases I and II. Specifically, for compliance purposes the NTG ratios for Phase III programs continues 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 amount 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, 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, 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

Under the Uniform Methods Project (UMP) funded by DOE, The Cadmus Group and its subcontractors have developed a framework and a set of protocols for determining the energy savings from specific energy efficiency measures and programs. The Phase I report, published in April 2013, outlines methods for evaluating gross energy savings for common residential and commercial measures offered in ratepayer-funded initiatives in the

United States.[134] Phase II addressed cross-cutting issues, including a protocol for determining NTG, published in September 2014. However, because definitions of net savings (for example, whether it includes participant and/or nonparticipant spillover) and policies regarding NTG vary across jurisdictions, the UMP net savings protocol does not offer specific recommendations on how NTG is applied nor does it offer specific survey questions and analysis techniques.

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 memo.

The *Framework* also defines participant and nonparticipant spillover and recommends the consideration of trade ally surveys and reports for assessing the nonparticipant portion of a programs 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 III. This method applies to downstream programs, typically using some incentive or direct installation.[135] Research Into Action and Energy Trust of Oregon developed this approach for telephone and on-site assessment of NTG (by project and by measure) across residential, commercial, industrial, and government sectors including:

- Rebates and grants for energy efficiency improvements
- Rebates and grants for renewable energy sources

[134] The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures, January 2012 – March 2013, NREL/SR-7A30-53827 published April 2013 by The Cadmus Group, contract no. DE-AC-08GO28308 and found at http://www.nrel.gov/docs/fy13osti/53827.pdf

[135] At the November 2013 PEG meeting, the SWE offered a memo on NTG approach for Appliance Recycling Programs based on the Uniform Method Project. Finally, 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.

- 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 FR 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.[136]

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 evaluation contractors 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 that best describe 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.

The offered response options (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: a) the level of savings that the respondent would have achieved without the program’s assistance; and b) in the case of nonresidential programs, whether the respondent’s business or organization would have paid the entire cost of the equipment replacement or upgrade without the program

[136] See Section 2.4.4 for detailed discussion.

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.[137]

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.
- Challenging a respondent's consistency can make the respondent feel defensive and may not produce more accurate data[138] – 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 nonresidential and residential programs and how it can be further tailored if needed.

C.4.2 Assessment of Intention in Nonresidential Programs

In this section, we describe how the Common Method typically is applied and scored in standard, nonresidential incentive programs. We also discuss tailoring or modification of the Common Method.

General Application of Intention Assessment in Nonresidential Programs
Typically, the nonresidential battery begins with the following question:

[137] Enders, C.K. *Applied Missing Data Analysis*, New York: The Guilford Press, 2010.
[138] See Section C.6.5, *Incorporation of Trade Ally Responses* for a more detailed discussion.

- Which of the following is most likely what would have happened if you had not received [the program assistance]?

The battery has included the following options in multiple evaluations of a wide range of nonresidential 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:

- 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

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 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.

In response to the initial draft of this memo, some evaluators have noted 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. This is understood, but 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:

- Would your business have paid the entire cost of the upgrade?

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.[139] Respondents get the highest free rider score *only* if they report that they would have done the same project

[139] See Section C.6.1, *Controlling for "Socially Acceptable" Response Bias*, for a more complete discussion of this potential bias.

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 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 Nonresidential 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).[140]

Table 47 summarizes the possible response combinations to the questions described above and the *intention* score assigned to each unique combination.

[140] *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.

Table 47: 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**

* Represents the midpoint of possible values for this question.
** Infrequent response.

Tailoring of Intention Assessment in Nonresidential 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:

> *"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 LEDs:

1. Put off replacing the [X type of] lights with LEDs for at least one year or cancelled it altogether.
2. Kept some of the existing lights and replaced some lights with LEDs.
3. Installed different lights. If so, what kind? _____
4. Installed the same number and type of LED lights 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 lights would they have replaced with LEDs? 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

In response to the initial draft of this memo, some additional types of modifications have been suggested:

- 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 nonresidential 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 nonresidential 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.

Table 48 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 .25; and reporting the same outcome as the incented upgrade or purchase results in a score of .5.

Table 48: 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 nonresidential 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 nonresidential assessments.

C.4.4 Influence (Nonresidential and Residential)

Assessing program influence is the same for nonresidential 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. Among 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.

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.

Table 49: 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 50.

Table 50: 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 .01 to convert it into a proportion for application to gross savings values.

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 (DI) Program

Direct install (DI) 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 DI program, the battery sought to 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 was 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 was as described above, but included potential program influences reflecting the unique elements of the DI program. For example, in a case where the program included a building assessment along with DI measures, the influence question included "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 (EPC)

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 evaluators surveyed both the building owner and the ESCO.

The building owner instrument included the standard *intention* question of what would have happened (postpone/cancel, smaller project, same upgrade) without program support and the standard "influence" question.[141] The evaluators calculated building owner *intention* and *influence* following the standard approach, described above.

The instrument for ESCOs asked:

- 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 evaluators calculated only ESCO intention, using the algorithm shown in Table 51.

Table 51: 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

[141] Influencers were program information, interaction with program staff, the list of prequalified ESCOs, and program assistance in selecting an ESCO.

To aid in determining how to combine the building owner and ESCO scores, the building owner instrument also asked:

- Whether they had ever worked with an ESCO before.
- Whether they would have used an ESCO without program assistance.

The evaluators used the algorithm shown in Table 52 to calculate the intention component score based on responses by both the building owner and the ESCO. The algorithm assumed that the ESCO responses were not relevant if: 1) the building owner was experienced with ESCOs and so could accurate predict what would have happened without the program assistance; 2) the owner indicated that without program assistance they would have cancelled or postponed the project or would not have used an ESCO.

Table 52: 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 was also considered and could be used to adjust the building owner's score.

C.6 Response to Questions and Concerns Raised About the Common Method

In response to the initial and revised drafts of this document, some evaluators raised questions or concerns concerning the Common Method described above. We have revised the above sections to address those concerns. 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. We analyzed responses to the battery from 158 participants in nonresidential retrofit and new construction programs and 1,252 participants in a range of residential programs (appliances, shell measures, home performance, and refrigerator recycling).[142] First, the two components positively correlated in both the nonresidential 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 nonresidential (95% CI: 16.8 ± 3.4 vs. 5.3 ± 1.5) and residential (95% CI: 26.4 ± 1.3 vs. 10.5 ± 0.8) samples. If the shared variance 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.

[142] 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: Nonresidential & Residential Program Portfolio*. Submitted to Energy Trust of Oregon by Research Into Action, Inc., December 31, 2010.

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 nonresidential 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 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* counterfactual 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.[143]

We could not use the nonresidential data described above to reliably investigate the question of whether "don't know" responders differ from others, as only three nonresidential 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.[144]

We 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 results. As an alternative approach, we recommend using linear regression to predict the *intention* score from each respondent's *influence* score.

[143] Enders, C.K. *Applied Missing Data Analysis*, New York: The Guilford Press, 2010.
[144] The percentage of respondents who gave "don't know" responses to the *influence* component was even lower – 1% for both residential and nonresidential samples. Similarly, in a dataset of 228 nonresidential 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.

C.6.4 Consistency Checks and Related Issue

Consistency checks are frequently used in social and epidemiological research, but a Google search found many more references to using consistency checks to aid data cleaning after survey completion than to resolve seemingly inconsistent response while the survey is ongoing. 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.[145]

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 Incorporation of Trade Ally Responses

One evaluator asked how an algorithm for a residential program might incorporate trade ally responses in a manner similar to the ESCO example given in Section C.4.2, above.

The evaluator may propose an approach for SWE review (see Section C.6.2).

C.6.6 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.

[145] 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 net-to-gross ratio (NTG) for Phase III of Act 129 is to be treated in the same way as for Phases I and II. Specifically, for compliance purposes the NTG ratios for Phase III II programs continues 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 nonparticipant spillover ("spillover" or "SO") and recommends the consideration of trade ally surveys and reports for assessing the nonparticipant portion of a program's spillover impact. However, the SWE has determined that while estimation of nonparticipant spillover is desirable, it is not required. If assessed, nonparticipant spillover may be assessed through either a general population (nonparticipant) 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 assessment of participant and non-participant spillover.

As is the case with the common approach to free ridership estimation, EDCs and their evaluation contractors 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 evaluation contractors 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 nonparticipant 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 nonparticipant spillover is desirable, it is not required. If assessed, the sampling approach should produce a sample

that is representative of the target population (nonparticipants 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 nonresidential 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 nonresidential participant surveys are slightly different.

D.3.1 Overview of Recommended Common Protocol

For both the residential and nonresidential sectors, the participant spillover approach will assess, 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, below.
- An estimate of energy savings associated with those energy efficiency measures. (Details in Section D.3.2, below.)
- 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, .5, and 1. (Details in Section D.3.2, below.)

The specific methods for the residential and nonresidential 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:

- 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?[146]
- 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 program influence for different measures. At the evaluator's discretion, the survey may assess influence for each measure identified.

[146] 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.

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)?[147]
    - 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 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, a working group of the PEG will develop conservative working assumptions for any required inputs (e.g., square footage of home, R-value improvement, replaced wattage). As an alternative, the PEG working group may identify average verified savings for such measures.

[147] 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.

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 influence score:

$$Measure\ SO = Measure\ Savings * Number\ of\ Units * Program\ Influence$$

For each of the above categories, the evaluators then will:

- 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.

$$\Sigma Participant\ SO\ (population) = \frac{\sum Participant\ SO\ (sample)}{Sample\ n} \text{ x Population } N$$

- Divide that total savings by the total program savings to yield a participant spillover percentage:

$$\%\ Participant\ SO\ = \frac{\sum Participant\ SO\ (population)}{Program\ Savings} \text{ x } 100$$

D.3.3 Nonresidential 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 nonresidential participant spillover is similar to that for residential, but differs in some details.

Identification of Non-Rebated Nonresidential 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 Nonresidential 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 2 or 3 = 0.5 (half of the savings attributed to the program).
- A rating of 0 or 1 = 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 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, a working group of the PEG will develop conservative working assumptions for any required inputs (e.g., square footage of home, R-value improvement, replaced wattage). As an alternative, the PEG working group 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 Nonresidential Spillover and Savings Rate

The calculation of nonresidential 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 influence score:

$$Measure\ SO = Measure\ Savings * Number\ of\ Units * Program\ Influence$$

For each of the above categories, the evaluators then will:

- 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 NONPARTICIPANT AND TOTAL SPILLOVER

The SWE has determined that while estimation of nonparticipant spillover is desirable, it is not required. Nonparticipant spillover may be assessed either through a general population (nonparticipant) survey or through a survey of trade allies.

D.4.1 Nonparticipant Survey

If a general population survey is selected, it should assess, 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 participant's decision to take the identified measures, assessed with a rating scale and converted to a proportion, with possible values of 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 nonresidential participant surveys are slightly different and are described in separate subsections. The residential and nonresidential 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 nonparticipant spillover through trade ally surveys, separate surveys should be conducted for the residential and nonresidential sectors. Each survey should assess, 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 TAs 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 (details in Section, below).Trade ally survey questions should ask about sales within a specific sector (residential or nonresidential). 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 commonly 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 commonly 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 nonresidential) and to all of 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 similar to 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 commonly sold program-rebated measures, the number of non-rebated units will be estimated as 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:

- 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 similar to 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, a working group of the PEG will develop conservative working assumptions for any required inputs (e.g., square footage of home, R-value improvement, replaced wattage). As an alternative, the PEG working group 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:

$$Reported\ Measure\ SO = Measure\ Savings * Number\ of\ Units * Program\ Influence$$

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 Nonparticipant 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 nonparticipants.

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 nonparticipants by subtracting the total participant SO for that sector from the total trade-ally-reported SO, as:

$$\sum Nonparticipant\ SO = \sum Total\ TA\ Reported\ SO - \sum Participant\ SO$$

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 nonparticipant SO. In that case, evaluators should report the total trade-ally-reported SO and participant SO.