New Jersey Evaluation Guidelines:

Enhanced Rigor Process and Impact Evaluations for On-Going and Mature Programs

May 2023

Guidelines for Clean Energy Program Evaluations
Sector: Residential and Commercial Evaluation Studies – On-going and Mature Programs
Evaluation Type: Process and Impact Evaluations

Prepared by Statewide Evaluators (SWE) as part of Assignments for NJ BPU

Prepared for:
New Jersey BPU Staff and NJCT Committee
Client Contact: Philip Chao

Final Document
May 22, 2023,
Replacing 2/16/22 version

Contents

Abstract

This document focuses on guidance for program evaluation efforts by utility and state Independent Program Evaluators (IPEs) for residential, multifamily, and commercial programs that are on-going or mature. For the impact evaluation elements, it covers guidelines for evaluation of programs that have a year of pre- and post-data available under similar design or delivery. These guidelines do not cover new or transitioning programs, and do not cover behavioral programs. Evaluations of those programs are covered by separate Guidelines documents. Note that the term "program" in this document is intended to cover traditionally understood programs, and may apply to programmatic efforts that utilities or others may sometimes refer to as "sub-programs" or other terms. If a question arises about what is intended to be covered by these guidelines, the SWE will clarify as needed for each instance. Note that this document includes instructions for:

- Minimum requirements for the work plan, and an associated summary table to be included with each project's work plan;
- Minimum expectations for data collection, analysis methods, and statistical rigor associated with process and impact evaluation efforts;
- Minimum expectations for evaluability assessment efforts;
- Major stages of working arrangements with the SWE; and
- Minimum expectations for outputs from the study, and the content and timeline for the report.

Other NJ Guidelines are available for new and transitioning programs, and other specific types of evaluations (NTG, behavioral, others).

There are expectations of working with SWE in the preparation of the evaluations conducted in association with these Guidelines, including:

- Review of scopes for conformance
- Regular meetings to monitor progress related to the study and conformance
- Review of key items including sampling plans, survey instruments, data collection methods, analytical methods, and similar for conformance
- Discussion of draft analysis results and findings
- Review of draft and final reports for conformance.

Determining Type of Evaluation Study Required

Table 1: Summary of Evaluation Study Expectations

	Process	Impact	Notes
Basic Guidelines	One (or more) per year, as long as the program remains "new" or changing	One per year, as long as it remains "new" or changing	No program should be "basic" for 2 years without discussion with SWE. Most are 1 year maximum.
Enhanced Guidelines, before and during Tri2	Minimum 2 per triennium per program	Minimum 2 per Triennium; may be 1 if program is well-established and is low percent of savings.	Need robust NJ data for TRM; lighting going away and need updated numbers and values for "newer" measures that will increasingly be the

	Process	Impact	Notes
			core of programs; most programs did not get strong-sample process evaluations in completed first-year evaluations.
Enhanced Guidelines, after Tri2	Minimum 1 per Triennium	Minimum 1 per Triennium unless PJM has more frequent requirements	Mature programs and TRM values will be more settled. This keeps up with some of the program changes.
Behavioral	Annual, unless discussed with SWE	Annual, unless discussed with SWE	It is assumed that the randomized control group is arranged and evaluations are straightforward.
Net-To-Gross	Prefer 1 (or more) for each program and key measures / end uses in a Triennium for all high-priority, high-savings programs. If not conducted at the utility level, Integrated with Basic or Enhanced rigor surveys, the State will conduct the studies.		

Study Delivery Timing:

Studies do not have to be in synch with program years (PYs); however, except for perhaps first year basic guideline process and imipact work, which can be conducted on data that is not a full year, the studies should be based on at least 12 consecutive months of data. It may represent 6 months of one program year and 6 months of another, or other configurations that work with efficient evaluations and data availability.

Delivery of the final evaluation studies prior to the deadline for the Evalution Use memo and the next annual or comprehensive update to the TRM (December 1) are expected. Completion prior to preparation of Annual Reports tracking is strongly encouraged (mid-September). For basic studies on new programs, the fastest turnaround possible after data collection is preferred, so the recommendations can be implemented quickly and programs "righted" as may be needed, and the effectiveness of the changes can be verifed through the next rapid-turnaround basic or enhanced evaluation work. Planned schedules wil be reviewed with SWE.

1.0 Introduction

The Enhanced Rigor Evaluation level applies to mature programs with well-established design, delivery, administration and participation rates, and at least one year of performance data.

Program evaluations will inform the New Jersey Investor-Owned Utilities (Utilities), the State, the Board of Public Utilities (BPU) collectively the program administrators (PAs) and the SWE on the operation and functioning of the Energy Efficiency programs operated within the current and future New Jersey Clean Energy Programs (NJCEP) portfolio. This document covers programs managed by the utilities and the State.

Evaluation is not a report card, it is a management tool, and it is important that the Utilities and BPU (the NJCEP PAs) should focus primarily on improving the design and execution of the programs, the *a priori* savings estimates for the next planning cycle, and targets for the next Triennium.

These guidelines are a set of study requirements to help focus the Independent Program Evaluators (IPEs) and State Independent Evaluation Contractors (IECs) of NJCEP PA managed programs on how studies should be designed and implemented. Following this guidance:

- All NJCEP PAs (state and utilities) and their evaluators will use these guidelines to create an evaluation work plan for each program study and will submit the draft work plans to the SWE.
- The SWE will review these work plans as expeditiously as possible and, after review, comment, and discussion with the parties, approve final work plans.

The SWE understands that there will be specific circumstances where alternative approaches may be appropriate. An NJCEP PA and its IPE or the State IEC may submit alternative approaches along with a detailed explanation of the approach and an explanation as to why this approach is appropriate and feasible for that study.

These guidelines are not intended to limit the scope of the evaluation; they are minimum expectations related to the impact and process elements of evaluation studies of NJCEP Energy Efficiency programs.

Note that NJCEP utility program administrators may, and are encouraged to, provide joint work plans for individual programs, as this would be expected to lead to evaluation cost savings However, the plans should include sampling plans that meet the recommended sampling precision (10% at 90% confidence) for each program, and 90/15 or 90/10 for significant measures.[1] SWE will discuss with the State evaluators whether the study should provide the required precision at the state level, or at the utility territory level. In addition, for the sake of evaluation economy, the guidelines include a recommendation for the studies to collect data to support NTG studies.

2.0 Residential Impact Evaluation – Enhanced Rigor

The impact evaluation is designed to provide reliable information and estimates for the TRM and program guidance on:

- gross and net savings by fuel type and key measures[2];
- realization rates; and
- causes of and actionable recommendations to improve realization rates (where possible).

For studies with time and funding, and in consultation with the SWE, additional auxiliary / complementary evaluation efforts may also be included, for example:

- Retention and persistence of measures, to track the retention of installed measures.
- Customer profiling and equity assessments, to review how past and recent participation rates vary across the state, normalized against sets of common values like social, housing, economic, demographics, and other census-available information.[3]
- Targeting via census feasibility, to leverage off the information gleaned above; and

[1] The SWE anticipates that since the evaluation should address all significant measures, the overall sample size program wide will support better than 90/10 in most cases.

[2] By "key measures" we mean measures that represent a total of at least 80% of the program's savings and including at least 3 measures that are not lighting. If the program is a pure lighting program, this is not applied, and if there are fewer than 3 other measures, or there is a case for why 3 other measures are not appropriate, address this in the workplan.

[3] Include a plan to address issues that may be important, like self-selection bias, etc.

- Targeted review of new or emerging, changing, high impact measures, installed by multiple programs.
- Modules for non-energy impact assessments, occasionally, to leverage the surveys used for impact or process evaluations.
- Specific parametric studies to address other important analytical issue or leverage data opportunities.

The results of these elements may be included in the TRM, but are also suited to program design, progress monitoring, and delivery guidance applications.

2.1 Data and Methods

The enhanced rigor impact evaluation will use data gathered from participant surveys (e.g., household characteristics, installed measures, etc.) and company records (e.g., program records including measures installed, and pre-post billing data. Any portion of the work that involves sampling must be described (including the participant survey), and the sampling design provided. The sampling design should focus on describing the strata, rationale for the stratification plan, and where possible, have orders-of-magnitude expectations for goal responses for each reporting metric The SWE recommends that the evaluation include stratification and sample size sufficient to provide information on measures that represent a total of at least 80% of the program savings at the program level, and must include any measures representing more than 5% of the program savings, and that at least two measures beyond lighting must be included (at the program level). Because evaluations are intended to be forward-looking, measures that are expected to increase to more than 5% of savings in the next period should be included. The SWE recommends total end-of-year sample sizes should provide at least +/- 10% at 90% confidence overall at the program level, and 90/15 for specific measures or targeted subgroups/strata at the program level for each utility (90/10 if the number of sample points in the subgroup is 1,000 or more).[4]

To achieve the goals of gross and net savings by fuel type and measure, one or more of the following three analytical impact evaluation approaches are generally expected to be employed, depending on the characteristics / types of measures in the program.[5] Specialized or targeted impact research such as hours of use or load shape studies may use additional approaches. In all approaches data preparation steps should be described in the report, and the loss of sample due to any checks or steps must be detailed and the implications and concerns discussed.

- **Billing analysis**: in conjunction with the engineering analysis, the billing analysis of participants estimates the savings associated with subsets of key program measures. This approach is robust, but often cannot address measures representing small amounts of savings, or individual measures installed as a "bundle". The billing analysis will:
    - Use a monthly post program regression billing analysis model, or similar (with justification).
    - Include explanatory variables for program measures, as well as to control for the impact of measures installed through other electric or gas utility programs or other sources;

[4] If the population is too small for these requirements, a solution should be discussed with the SWE.
[5] If the number of program measures is small, or the measures do not meet some of the conditions noted, the case may be made for using one or two methods, but the explanation must be included in the scope.

    - Uses a control group of "future" participants, identified using a matched control group, participant scoring, or other similar technique.
    - Results, coefficients, confidence intervals, and other documentation should be provided.
- **Building simulation**: Estimates of gross measure-specific energy savings for measures with known or potential interactive effects for which billing analysis is not feasible, may be generated using building simulation modeling. Suitable software is available from NREL, DOE, and other sources.[6] The model can be calibrated using billing data results for heating, cooling, and baseline usage. The final set of models and workbooks should be available to document the work.
- **Engineering algorithms**: Engineering algorithms are used to estimate energy savings for measures not well estimated using billing analysis (e.g., small savings, bundles, etc.), and measures not impacted by interactive effects. The priority of suitable algorithm sources would include New Jersey, nearby states, well-respected national sources. The computation and inputs should be documented in a workbook available to the SWE.

In addition, work scopes should include estimations of realization rates and NTG data collection / analysis and verified in-service rates.

- **Realization Rate**: The pre- and post-consumption for each participant should be computed and compared with the a priori estimates. Again, the results should be computed by measure, end-use, and program wide. The evaluation would be expected to look for patterns and disparities by dwelling type, building size, climate zone, insulation, household characteristics, vendor, or other potential causal factors. Confidence intervals must be reported.
- **Net to Gross**: These Guidelines also expect a NTG analysis and/or the NTG data collected and provided to the State IEC for computation, using specified / adapted survey questions and analytical methods as described in the Adopted NTG Guidelines. The efforts should provide (or gather data to support) NTG by measures, end-use, and program-wide to the degree possible. Stakeholder interviews may also be used to provide additional shading on the NTG results. Confidence intervals must be reported, and the results for free ridership should be reported separately from spillover, and the combined value also reported. Freeridership, spillover, and combined values will be developed by the State IEC and included in the TRM; PA IPEs may report additional NTG findings for their service territory.
- **In-Service Rates:** Observation of a sample of customers at 90% confidence +/-10% precision should be used to estimate in-service rates. If this is infeasible, self-report using interview or survey data may be used.

PJM: If the peak demand savings from the program and/or measures is offered for sale on to PJM, the evaluation work plan must include specific detail on the approach planned for meeting the rigor and documentation required to meet PJM standard.

Results are expected to provide gross and net savings and realization rates by fuel type, end use, and measures; as well as data or results for NTG at the same level of granularity. These values are expected to for use in the TRM update.

[6] Including Beopt, Energy Plus, etc.

3.0 Commercial and Industrial (Non- Residential) Impact Evaluations, Enhanced Rigor

The Enhanced Rigor Impact evaluation is designed to provide reliable information and estimates for the TRM and program guidance on:

- gross and net savings by fuel type and key measures[7];
- realization rates with explanations for significant (±10%) deviation from 100%;
- prospective realization rates for programs significantly impacted by approved TRM updates or as advised by the SWE;
- comparisons to similar programs in similar jurisdictions;
- reliability and accuracy of program tracking and ex ante savings calculations methods; and
- where possible, reasons for, and actionable information on remedies to improve the realization rates.

Following is some general guidance for enhanced rigor impact evaluation plans for each program.

- Enhanced rigor impact studies incorporate basic verification with site-specific measurements to collect energy performance data used in the engineering analysis of a measure's baseline and post-installation performance. Site-specific information can be collected through physical or virtual site visits, participant surveys or interviews, facility billing or other consumption data. For prescriptive projects the required data are the independent variables defined in TRM algorithms. For custom projects the data are the independent variables described in site-specific measurement and verification plans (SSMVP) for installed measures.
- Sampling of projects by program is expected in the commercial & industrial (non-residential) sector. The SWE recommends that the evaluation include stratification and sample size sufficient to provide information on measures that represent a total of at least 80% of the program savings at the program level, include measures representing more than 5% of the program savings. At least two measures beyond lighting must be included (at the program level). Because evaluations are intended to be forward-looking, if there are measures that are expected to increase to more than 5% of savings in the next period, then these measures should also be included. The SWE recommends total end-of-year sample sizes should provide at least +/- 10% at 90% confidence overall at the program level, and 90/15 for specific measures or targeted subgroups/strata at the program level for each utility (90/10 if the number of sample points in the subgroup is 1,000 or more).
    - Site visit measure level sampling can be used when verifying high measure counts would result in excessive labor cost. Precision for site visit sampling is ±20% at the 90% confidence level.
    - Sample plans should explain any deviation from these confidence/precision guidelines.
- Measurement & verification (M&V) plans for key measures are required for each plan. M&V plans should reference an International Performance Measurement and Verification Protocol (IPMVP) option (A, B, C, D), describe the analytical approach, expected engineering model(s), reporting metrics and associated uncertainty.

[7] By "key measures" we mean measures that represent a total of at least 80% of the program's savings and including at least 3 measures that are not lighting. If the program is a pure lighting program, this is not applied, and if there are fewer than 3 other measures, or there is a case for why 3 other measures are not appropriate, address this in the workplan.

For studies with time and funding, additional auxiliary / complementary evaluation efforts may also be included, for example:

- Persistence of measures, to track the retention of installed measures;
- Customer profiling and equity assessments, to review how past and recent participation rates vary across the state, normalized against sets of common values like social, housing, NAICS code, economic, demographics, consumption, and other census- or company-available information;[8]
- Targeting via census feasibility, to leverage off the information gleaned above;
- Targeted review of new or emerging, changing, high impact measures, installed by multiple programs;
- Modules for non-energy impact assessments, occasionally, to leverage the surveys used for impact or process evaluations; and
- Specific parametric studies to address other important analytical issue or leverage data opportunities.

The results of these elements may be included in the TRM, but are also suited to program design, progress monitoring, and delivery guidance applications.

3.1 Data and Methods

The enhanced rigor impact evaluation will use data gathered from program tracking records, project files, consumption records, and information gathered during site visits or participant surveys. Project and measure sampling is expected must be described (including the participant survey) in the evaluation work plan or survey design document, and the sampling design provided. The sampling design should focus on describing the strata, rationale for the stratification plan.

To calculate the ex-post gross and net savings by fuel type and measure, the following steps are expected:

- **Review program tracking data and calculations:** This quality control step:
    - compares fields in the program tracking system to those needed or expected for the program type;
    - checks that all fields are populated, the units (e.g., kWh, tons) are correct; and
    - confirms or not that interim and savings calculations use correct TRM algorithms (except custom).
- **Desk review:** conducted on the sampled project files:
    - compares project file data (dates, counts, ratings, incentives, etc.) to the program tracking record;
    - checks that each project file is complete, supports TRM or custom algorithms; and
    - includes sufficient information, and data if applicable, to establish measure baseline conditions.
- **Engineering review:** conducted on the sampled project files (as needed):
    - Uses simple engineering models (SEM) to independently check tracking system and/or TRM calculation; and
    - Generally, applies only to measures where tracking system savings are unclear, use novel methods, yield unexpected results.

[8] Include a plan to address issues that may be important, like self-selection bias, etc.

- **Verification:** based on site visits, project files and participant interviews; confirm that:
    - measures are still installed at the time of verification and performing as planned in the project file; and
    - installed measure counts, ratings, models, loads match values in the tracking data.
- **Site specific measured data** conducted on the sampled project files (as needed)**:**
    - Spot measure, log, trend using an energy management (EMS) system, the points needed to carry out the M&V plan for the program evaluation. Measurements should capture the measure's expected range of operating conditions. Measurement uncertainty should be within commonly accepted best practice limits.
    - Measured data can include EMS trend logs for both pre- and post-installation periods. Pre-installation data to establish baseline conditions should be pursued including from vendors and installation contractors.
- **Analysis:** one or more of the following approaches is expected in the impact M&V plan:
    - Simple engineering model using key parameter measurement (IPMVP Option A)
    - Regression analysis using end-use device/system metered data (IPMVP Option B)
        - The regression model, and raw and processed data must be included in the M&V documentation. The model must make physical and engineering sense. Outliers and discarded data should be preserved and explained. Uncertainty analysis must be included in the M&V results.
    - Regression analysis using consumption and weather/production data (IPMVP Option C)
        - The regression model, and raw and processed data must be included in the M&V documentation. Uncertainty analysis must be included in the M&V results.
    - Building simulation model (IPMVP Option D) based on as-built conditions.
        - The simulation program must be commonly accepted by evaluation professionals, transparent, publicly available. The model must be calibrated using consumption and weather data. Sources of model uncertainty should be identified.

In addition, the results should include realization rates, net-to-gross values, and verified in-service rates.

- **Realization Rate:** The results should be computed by measure, end-use, and program wide. The evaluation would be expected to look for patterns and disparities by business type, size, climate zone, building and equipment vintage, vendor, or other potential causal factors. Confidence intervals must be reported.
- **Net to Gross**: These Guidelines also expect a NTG analysis and/or the NTG data collected and provided to State IEC for computation, using specified / adapted survey questions and analytical methods as described in the Adopted NTG Guidelines. The efforts should provide (or gather data to support) NTG by measures, end-use, and program-wide to the degree possible. Stakeholder interviews may also be used to provide additional shading on the NTG results. Confidence intervals must be reported, and the results for free ridership should be reported separately from spillover, and the combined value also reported.
- **In-Service Rates:** Observation of a sample of customers at the 90% confidence +/-10% precision should be used to estimate in-service rates. If this is infeasible, self-report using interview or survey data may be used.

PJM: If the demand savings from the program and or measures is offered for sale on PJM, the scope must include specific detail on the approach planned for meeting the rigor and documentation required to meet PJM standard.

Results are expected to provide gross and net savings and realization rates by fuel type, end use, and measures; as well as data or results for NTG at the same level of granularity. These values are expected to be delivered for use in the TRM update.

3.2 Additional Analyses Beyond Savings, ISR, RR

These guidelines describe expectations but are not intended to limit the study.

Peak Analysis: Certainly, the impacts on peak demand reduction are also of interest. If the utility has AMI or other data or approaches suitable for examining the impacts of the program on demand, that should be included the scope and discussed with SWE.

NEB / NEI Analysis: Non-Energy benefits (NEBs) / Non-energy impacts (NEIs) analyses may also be incorporated into the evaluation to identify NJ-program-based effects. This may include NEBs/NEIs that are literature-based, survey-based, financial computations or other methods. Again, these may be included in the scope and the specific methods discussed with SWE.

4.0 Process Evaluation All Sectors – Enhanced Rigor

These evaluations should use the current year of program operation as the focus and be forward looking to see how the program might be able to improve and grow over the next two-to-three years. It is imperative that this guidance should not be construed as limiting the investment in added process and market evaluations. For instance, if a program over or underachieves goals or if new products and services are becoming available added process evaluations are encouraged. The process evaluation should also support the impact evaluation efforts.

Process evaluations include:

- a qualitative / quasi-quantitative review of the program's delivery, performance, and documentation, with a focus on actionable recommendations for program improvements, and in-depth information on program barriers and remedies.
- a quantitative data-driven review of program tracking data, reflecting program timing and backlogs at key stages; participation and uptake patterns; useful comparisons of contractor performance; and, potentially, equity, access and participation patterns by sensitive customer groups.

4.1 Process Evaluation Study Questions

The study questions will vary by the type of program and the delivery methods used for the program. The expectation is that the evaluation team will customize each process evaluation for a program based on the utility's experiences with the program. There are four domains to consider when developing research questions:

- Program Administration – this domain concerns the various administrative processes that support the program, back-office activities the utility and the implementation contractor use to support the program and the various processes the customer or trade ally must follow to gain approval of the project and receive any incentive or other support. A thorough review of program materials, tracking data, and outreach materials to review consistency and effectiveness should be included.

- Program Implementation and Delivery – this domain concerns the interactions with customers and trade allies and other stakeholders, this includes the marketing process, meetings with customers, the delivery of services, how retailers present products, how manufacturers market, quality of delivery, how quality is assured, detailed review of barriers, etc.

- Market Response – this domain concerns the market and how those market actors, trade allies, stakeholders, retailers, manufacturers, etc., respond to the program, their awareness of the program the products, the services, any barriers to the program, any changes in the market offerings resulting from the program, the use of the marketing materials, the effectiveness of the marketing efforts, whether the market structure changes or results in resistance to the program, satisfaction with the program, and what leads to adoption.

- Customer Response – this domain concerns the utility customer and their experience of the program, important is analysis of the characteristics of participants and comparison to nonparticipants (as defined in the approved work plan) and to program goals, also whether the program addresses barriers to adoption, how the customers learn of the program, product or service, satisfaction with the program delivery, or service or product, what leads to adoption or rejection of the product or service, and use of the product or service following program participation.

The development of a study plan should begin with the evaluation team interviewing the program implementation staff (both utility and implementation contractor, including the marketing team) to assess what type of information these people need in order to grow and evolve their program over the next two years given the constantly surfacing set of new products, regulations, and changing market conditions. The four domains should provide guidance that all aspects of the program are worthy of investigation, but the final study plan should focus on the highest priority research areas to ensure the ongoing effectiveness of the program.

4.2 Sampling and Data Collection – Process Evaluation

To address the study questions the evaluator should interview program staff and implementation staff, and survey samples of participants, partial participants (i.e., customers who take some level of action to participate in the program but do not complete the process) as applicable, and trade allies who support the program. The SWE recommends that the samples should be drawn quarterly, and high-level results

reviewed quarterly, to supply feedback and allow problems to be addressed promptly. If participation is too low to allow this frequency, other frequencies can be proposed.

The evaluations should provide at least +/- 10% at 90% confidence overall at the program level; and if addressing measure specific issues, 90/10 for measures accounting for at least 80% of the program savings, any specific targeted subgroups/strata, any measures representing more than 5% of the program savings and include at least two measures beyond lighting. Because evaluations are intended to be forward-looking, if there are measures that are expected to increase to more than 5% of savings in the next period, then these measures should also be included. Data collection for the process and impact evaluations can be combined for efficiency or may remain separate. Trade ally and staff interview sample sizes should be sufficient to provide input into relevant research objectives.

Depending on the size of the program, vendor IDIs or stakeholder interviews may include from 20-40 completions, or less for a very small program, or if the contractor provides a rationale for a different sample size. Participant surveys need to include stratification that supports a minimum of 90/10 for each key measure (see impact evaluation criteria for key measures); the overall program confidence will exceed 90/10. Besides measures, strata may also need to address low / moderate income participants for residential, or business (consumption) sizes or business type for commercial. Sample sizes may also be influenced by the numbers needed to support the impact evaluation work.

Survey and interview guides must use multiple survey items/questions to address the research questions and those questions should be specific to each program. The evaluator also should triangulate the primary data collection with data from the program tracking database, program collateral and a review of program participation forms and documentation prior to drawing conclusions. Multiple data sources should support conclusions and recommendations.

4.3 Data Collection Instrument Format

The SWE recommends that data collection instruments being submitted for review include the following information:

- Title: including contact type (e.g., program staff, participants, non-participants, partial-participants, trade allies, industry experts)
- Statement of purpose (summary for interviewer, client, and survey house)
- Listing and explanation of variables to be piped into the survey and the source (i.e., survey, database, etc.) of these values (if applicable)
- The topics that the scope identified as key for the survey, and the key outputs planned.
- Note that the SWE strongly prefers consistent questions across survey / interview groups within a program evaluation (with tailoring) for comparability.

5.0 Evaluability Assessment

Every first evaluation of a program is expected to include a specific evaluability assessment. The purpose of this activity is to provide early assurance that the data collection and data access can fully

support the needed process and impact evaluations expected of all EE programs in the portfolio for which savings are claimed. Early investigation is required so any necessary changes in data collection or procedures can be implemented prior to the next evaluation. The expectation is that the IPEs will verify that all variables needed from the program tracking data, from billing records, worksheets, and all other sources that will be needed to support an Enhanced-Rigor process and impact evaluation of the program are being collected, are populated, are accessible, and are accurate. The product of the evaluability assessment is a clear statement in the report that the IPE confirms they investigated and reviewed the variety of specific types and sources of data needed, and that the data were present, accurately collected, available, and populated. The confirmation statement should list the various types (not individual variables) that were verified, and that the IPE confirms that the data to support Enhanced Rigor Process and Impact evaluations of the program can be supported. If the evaluability assessment finds the data or processes are lacking, specific recommendations to remedy the issue(s) should be provided clearly and specifically in the report.

Note this evaluability assessment will need to be repeated in any evaluation in which the data collection, procedures, or other processes have changed that may affect aspects of the development of data needed to support Enhanced Rigor Process or Impact evaluations for the program. If no such changes have occurred, the IPE may cite and repeat the previous evaluability statement in the next evaluation. However a statement of evaluability must be included in each evaluation conducted on the programs.

6.0 Analysis Methods, Findings and Forward-Looking Recommendations Focus

Providing Context/Benchmarking: To support the evaluation recommendations, the reports should provide clear supporting findings from the research, and from comparisons of these findings with past research on the NJ programs as well as comparisons to other strong-performing similar programs in other locations. Therefore, each process and impact evaluation is required to include a chapter within the report summarizing key results from several other similar programs elsewhere. These other programs should provide benchmarking information that the NJ programs can refer to better put NJ results in context and potentially identify strong or better practices in the program type. Results from these programs should be referred to in multiple places in the report, noting where satisfaction, or savings, or other results are higher or lower than the ranges identified in other programs, or where they have improved or not improved compared to previous cohorts of the NJ program.

Analytical Methods and Clarity of Results: For the range of analyses conducted in the report, at least, the following methods and guidelines should be used:

- Results should be reported out in a way that allows straightforward comparison of results for specific subgroups (e.g., participants and partial participants in adjacent columns, etc.). Graphic results, including stacked bars to 100%, can illustrate results well. All relevant tables should include confidence intervals as well as the point estimate. Likert scales and Categorical responses: Percent reporting each categorical response and observation counts, and confidence intervals where appropriate.

- Labeled scaling: Percent reporting each categorial response and a weighted average and response counts, and confidence intervals where appropriate.
- Open End / Drill-down and Detail: Provide summary results using key words / intentions, and details as appropriate and meaningful / relevant for program changes going forward.
- Numeric responses: Means, averages, ranges, confidence intervals and response counts.
- Impacts: difference between reported installed vs. verified from surveys; effects on savings using TRM calculations, verifying the accuracy of implementation of the TRM steps. Response counts should be provided as appropriate.
- Models / regressions: as appropriate to attribute results to key factors. Supporting information should detail number of observations, confidence intervals for key outputs, etc.
- Comparisons of results: Comparisons over time within NJ, as available, and to similar programs in other states to illustrate trends, benchmarks, design/delivery/performance differences, and best practices. Comparisons should be made to programs that are as similar as possible; but even if identical programs are not available, lessons can be learned from comparisons to programs with similar elements. SWE assumes the independent evaluators have access to, and expertise in, such studies.

Required Results:

The goal is to provide findings, conclusions and recommendations that can reflect performance, but especially can provide real-time improvements and *forward-looking* recommendations related to:

- Program design and delivery.
- Program savings calculations and realization rates overall and by measure or measure group.
- Testing of the *a priori* computation of savings, and updating of TRM values where appropriate, to be used in subsequent triennial periods.
- Adequacy of the data to support the evaluation and recommendations for data improvements and data gaps related to evaluation.
- Recommendations related to program goals, measures, targeting for maximum impact, and recommendations for improvement to incentives, outreach, messenger, etc.

Impact results should focus on values to more accurately reflect program performance and update information included in the latest TRM. At a minimum, the enhanced rigor results should include:

- Tables of gross and net savings and realization rates by fuel type, end use, and measures; [9]as well as data or results for NTG at the same level of granularity.
- The State IEC will develop freeridership, spillover, and combined values, and report them in the TRM;
- IPEs may report additional NTG findings.
- IPE NTG research must be available for use for TRM updates.
- Other information gathered in the study that provides performance results by measure, measure group, or program-wide.

[9] Including all appropriate adjustment factors in the TRM

7.0 Reporting

The following guidance pertains to report format, reporting frequency, data collection and report delivery timing, and report and communication style.

7.1 Report Format

The following are requirements for all evaluation reports that will be submitted to the SWE.

The report should include the following:

- A 1–2-page abstract including list of all process and impact recommendations and clear tables of all the TRM update values including confidence intervals, observation counts, etc. (not just a list of what was investigated). This is separate from and in addition to the executive summary. The 1–3-page abstract briefly summarizes why the evaluation was conducted, and focuses on all quantitative results of any kind relevant for the TRM, and all program-related recommendations (without detailed explanation/context).[10] The evaluability confirmation and any related recommendations is provided in the Abstract.
- The Executive summary chapter includes more detail than the abstract. It clearly lays out results and recommendations with enough explanation and context enough to provide the reader with an understanding of the key elements and forward-looking results from the study. The evaluability confirmation and any related recommendations is provided in the Executive Summary. The Executive Summary provides enough description of underlying data collection and methods to give confidence in the results.
- A distinct chapter must be included in the body of the report that provides a summary of similar programs elsewhere and past results for NJ, if any. The chapter provides impact values and process / design / delivery comparisons for multiple similar programs elsewhere, and comparisons to impact and key process values from the program for prior years in New Jersey if available. These values should be used as a basis for best practices recommendations, trends in improving results, etc. The chapter and comparisons are required, but these results should also be referenced liberally elsewhere in the report as relevant, so that the reader can understand the context for the impact and process evaluation findings, and for recommended improvements.
- The report must also include a section that provides documentation of any data that are missing or needed in order to complete a standard impact or process evaluation as an assessment of the evaluability of the program going forward. Associated specific recommendations to address gaps should be included.
- It is required that all data purchased for the project becomes the property of or accessible to all other NJ evaluations.[11]

[10] The TRM-relevant results from the study are then considered and reviewed by the TRM committee and go through the TRM update process.

[11] Utilities should make every effort to include agreement in contracts for purchased data so that it can be shared to other New Jersey evaluation.

- For each evaluation project, several stages of data must be saved, with adequate documentation, and under properly compliant security. This includes at a minimum: initial data requests from the utilities; raw and cleaned, weighted survey or interview data; several stages of processed data; and final analytical data sets. These data must be held by either the IPE or utility in a secure location for a period of 5 years after the Triennium and be available upon request (and without charge) to the BPU and their consultants.

7.2 Report Frequency

Obviously, evaluation results should be as current as possible. For enhanced rigor evaluations, it is likely that impact evaluations will be conducted on an earlier PY, but process evaluations should be conducted on the most recent PY.

Evaluators can request a different reporting schedule, but the SWE asks that programs results be provided as close to the program period as possible, issued as completed for a program, without waiting to be included in a final portfolio report.

For impact evaluations, as some impact methods may require multiple years, this guidance does not prohibit this when approved by the SWE.

Process evaluation surveys (including for NTG) should be conducted quarterly to reduce hindsight bias. Shorter time periods may also be valid, in some instances, and this guidance does not prohibit such when approved by the SWE.

7.3 Report Timing, Data, and Data Collection

Special considerations for data issues include:

- Timing and schedules for data-driven impact or process evaluations might deviate from the prescribed schedule. A heat pump impact study requiring twelve months of metering will have a non-standard reporting date.
- Programs results should be provided as close to the program period as possible, without waiting to be included in a final portfolio report.
- Regression and simulation models including input/output workbooks used in an evaluation must be retained and available for review by the SWE and BPU.
- In all approaches data preparation steps should be described in the report, and the loss of input data due to any checks or steps must be detailed in the report appendix and the implications and concerns discussed. Where problematic losses might be remedied through changes in data collection or other methods, recommendations should be included in the report.
- Data acquired for evaluation studies must be retained and available to the BPU and their consultants for 5 years following study completion. PII must be removed from the data sets.
- If an IPE collects NTG data through surveys or other sources, they must be provided to the State IEC.

7.4 Report and Communication Style

Clear and concise communication is important. The following can help improve the style of reporting.

- The report body should begin with conclusions / recommendations, then summarize the associated supporting analysis for these results. It should not be organized in a historical fashion, documenting the order of work performed, or with results provided separately based on the source of the results. It should avoid walking the reader through all the data collection and analysis steps to get to the conclusion. The key audience includes users of the results, not other evaluators. Chapters should not be organized by "results of this primary data collection", "results of this primary data collection…". Appendices may use this approach.
    - Text style should favor bullets over pages of paragraphs. Remember the goal is to communicate results to users, who are not evaluators, but commonly need to be able to skim to glean their results of interest. Callouts and graphics of important findings / conclusions are encouraged.
    - Tables and graphics are important and desirable methods of conveying results. However, very long sets of tables (e.g., comprehensive survey results) should be moved to the appendix, and the body should focus on key results with implications for the programs. Complete results / tables / crosstabs of survey / data collection efforts and results should be included, generally in the appendix.
    - Bolding, underlining, subheadings, bullets encouraged when they help draw out conclusions.
    - Do not bury the lead. The first sentence of each paragraph should be the topic sentence. Avoid multiple clauses before the key point.
    - Tables / figures must be able to stand-alone because they are often extracted. This means table names must fully explain the contents, and table notes explain variables and abbreviations as needed. All Tables should include the n values and where appropriate, confidence intervals.
    - Survey sampling, stratification, sample sizes, and rationale must be described in the report, with accompanying tables and counts. CVs must be reported, along with statistical confidence and precision. These elements must be included to inform sample sizes and budgeting needs for future evaluations of the program. Detailed aspects of this information can be in the appendices. All survey instruments and interview guides must be included in the appendices.
    - Barriers should not be examined ONLY using Likert agree-disagree scales. The data collection work must include (open-ends that provide) details on the barrier and drill-down/follow-ups that include suggestions for remedies that would have addressed the barrier for the respondent group. For these first-Triennium studies, similar open-ended follow-ups should be considered for low-scoring elements of other process satisfaction questions.
    - Details on methodology should be provided *in appendices*; include description of phases of data cleaning and counts of the loss of sample from each of the various data cleaning steps.

8.0 Preparing the Work Plans / Scopes of Work

The SWE will review scopes of work for conformance with these overarching guidelines. The scopes should be a source of documentation of the evaluator's approach to the following topics.:

- How the objectives will be met, and research questions will be informed and analyzed.
- A section outlining special research issues or context for the specific program being evaluated.
- A list of the utility data needed to support the evaluation.
- The other programs or states that will be included in the program comparison section.

- Program start date, anticipated participants in the Program Year (PY), and rationale for conducting one vs. two evaluations in the first Triennium, if deviating from the two recommended.
- A sampling plan, including a table identifying the samples sizes overall and by each strata / subgroup, for each quarter and annually, and the expected precision / confidence for each group. The plan may pull fourth quarter respondents from the first, or first and second month of that quarter for timing reasons. Provide the rationale for the measure and other strata included.
- A data collection plan, including the data collection method for each group (Table 2 for the scope), and a table that identifies the key topics to be included for each survey / interview group (Table 3 for the scope)
- Clarity in mapping how each of the key research questions will be addressed (and potentially triangulated) in terms of both data collection and appropriate analysis approach.
- Detail regarding how the measure counts will be verified, and the steps anticipated to assure as collection of as accurate data as possible. Detail regarding how the calculations and factors will be verified.
- Risk elements associated with the scope, and methods to address those elements.
- Tasks with activities and deliverables, key milestones, a schedule, and a list of key staff.
- A specific section clearly laying out any deviations that are less rigorous than the expectations included in this guideline document, and the rationale.

The minimum Work Plan requirements for each program / study combination includes two pieces: 1) completion of the following table, and 2) preparation of an accompanying word document covering selected issues for the studies.

Required Table: Completion of the following Evaluation Studies Summary Table (Figure 1), meets most of the above requirements. The table may be provided for one program evaluation, or a table with multiple columns is provided for a scope or Plan for the portfolio of evaluations being conducted. In the latter case, separate tables may be provided for residential vs. commercial programs, or they may be combined. Each column in the table represents an individual residential or commercial program's evaluation study. A column should not combine programs or subprograms. A "study" associated with these guidelines may be a process evaluation or an impact evaluation or a combined impact and process evaluation – and may include elements related to NTG. The table may be provided in Excel or Word.

Required Separate Text: The evaluators must also provide, for each study identified in the table, a clear, succinct, word summary (not in the table) that contains:

- A discussion of the research objectives and research questions, with tailoring for each individual program's issues and needs,
- A sampling and survey plan table that specifically calls out each respondent groups across the top with the intended response number, and all key topics for the evaluation down the side, and clear checkmarks or other indication or explanation of the key topics to be addressed by each respondent group,
- A discussion of risks and how they will be addressed,
- A list of utility data to be requested,
- A succinct discussion of each task and how the analysis will be conducted,
- Detail on how the collection of accurate data will be assured, and
- A table of milestones and deliverables and dates.

This combination of text and tables is the minimum requirements for the workplan for each study.

Figure 1: Evaluation Studies Summary Table (Table 1 for Scope)

EACH COLUMN is a separate study. *Abbreviation "N"=Number of observations*	Program, PY & Study Name (sample answers) Comfort Partners PY 23, Impact & Process	Program, PY & Study name (example for a process-only study)	Next study / Study for Program 2
STUDY NUMBER	CP-23-1 or #1 or any numbering system	2	3
PROCESS EVALUATION			
Process & impact together?	Yes	No, process only	
Program Year	2	2	
Study Start / end date	7/23-12/23	7/23-….	
Solo or with other utilities (list)	Across all		
Rigor level	Enhanced		
# program participants expected	600		
Program's expected share of portfolio savings	10% of portfolio, 50% residential		
Types of Program materials to be reviewed (tracking, messaging, outreach, web, etc.)			
Staff, method (~N)	IDIs/ ~5		
Participant method, (order of magnitude N or precision/confidence),	Web Survey, stratified by measure, 95/5, combined with Impact		
Partial Participant method, (order of magnitude N or precision/confidence)	90/10, phone		
Non-Participant (order of magnitude N? or precision/confidence)	No		
Vendor / contractor surveys (N/precision), specify group / groups	Contractors, 30, phone, 85/15		
Measure or end uses? (specify key ones)	HVAC, Lighting, Wx		
NTG survey included? How many "N"?	Yes, 96		
NEI survey included? How many "N"?	Yes, abbreviated, 96		
Special research topics / research questions? *(Very important & tailored - Be sure to include detail in the Plan).*	Electrification		
Other notes, items included…			
Date and PY for last process evaluation	6/22-12/22, PY1	None	
Rigor level for last previous evaluation	Basic	None conducted	
Was evaluability resolved in last evaluation?	Yes	N/A	
States/utilities for comparison (included in body of report)	MA, MD, CT, CA		
IMPACT EVALUATION			
Process & impact together?	Yes	No	
Program Year	2		

EACH COLUMN is a separate study. *Abbreviation "N"=Number of observations*	Program, PY & Study Name (sample answers) Comfort Partners PY 23, Impact & Process	Program, PY & Study name (example for a process-only study)	Next study / Study for Program 2
Study Start / end date	7/23-12/23		
Solo or with other utilities (list)	Across all		
Rigor level	Enhanced		
# program participants expected	600		
Program's expected share of portfolio savings	10% of portfolio, 50% residential		
Staff, method,(~N)	IDI s, ~5		
Participant (order of magnitude N or precision/confidence), and survey method	95/5, web survey, combined with process Plus 30 on-sites		
Partial Participant (order of magnitude N or precision/confidence)	90/10, phone survey with process		
Non-Participant (order of magnitude N or precision/confidence)	No		
Vendor / contractor (N/precision), specify group / groups	No		
Measure or end uses? (specify key ones)	HVAC, Lighting, Wx		
In-service / verification planned? N, Method	Yes, >100 by phone survey		
Impact evaluation(s) method planned	Desk Review + billing analysis		
TRM generation applied	2022 Comprehensive		
NTG survey included? How many "N"?	Yes, 96		
NEI survey included? How many "N"?	Yes, abbreviated, 96		
Special research topics / research questions? *(very important/ tailored; be sure to include detail in the research plan)*	Small vs. large businesses / disadvantaged areas		
Other notes, items included…			
Date and PY for last process evaluation	6/22-12/22, PY1		
Rigor level for the previous evaluation	Basic		
Was evaluability certified in last evaluation?	Yes		
Other evaluation type			
States/utilities for comparison (included in body of report)	MA, MD, CT, CA		

9.0 References

Questions: Contact Jane Peters (janestrommepeters@outlook.com), Lisa Skumatz (skumatz@serainc.com), ore Dakers Gowan (dgowans@leftfork.com) - (SWE) .

The SWE considers the following documents as further guidance for New Jersey CEP Evaluations in general, these are not specific to New Jersey but many aspects of these apply such as definitions of rigor level, exclusive of specific state policy related content in the below documents:

a. California EM&V Protocols - http://calmac.org/publications/EvaluatorsProtocols_Final_AdoptedviaRuling_06-19-2006.pdf
b. California EM&V Framework - https://library.cee1.org/content/california-evaluation-framework
c. Pennsylvania EM&V Framework - https://www.puc.pa.gov/media/1584/swe-phaseiv_evaluation_framework071621.pdf
d. New York Process Evaluation Protocols - https://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/766a83dce56eca35852576da006d79a7/$FILE/Proc%20Eval%20Protocols-final-1-06-2012%20revised%204-5-2013.pdf

SWE anticipates these guidelines may be updated over time as needed.