New Jersey Evaluation Guidelines:

Basic Rigor Process and Impact Evaluations for New and Transitioning Programs

May 2023

Guidelines for Clean Energy Program Evaluations

Sector: Residential and Commercial Evaluation Studies – New and Transitioning 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 new or undergoing significant transition, including transitions in management, design, or delivery. These guidelines do not cover on-going programs that have a year of pre- and post-data available under similar design or delivery, 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 on-going and mature 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; and
- 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 core of programs; most programs did

	Process	Impact	Notes
			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 progamr 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 will be reviewed with SWE.

1.0 Introduction

The State of NJ offers a suite of Energy Efficiency programs operated by utilities, state / BPU, and others within the current and future New Jersey Clean Energy Programs (NJCEP) portfolio. Some are on-going, some periodically transition implementation between the BPU and utilities, and some are new programs developed by the state or utilities. Evaluation is used to identify the problems as well as the successes programs operating in the State. The SWE recognizes that newer programs are not functioning at full capacity during their learning periods. However, goals are in place and programs should be striving to achieve those goals.

Evaluation is not a report card, it is a management tool, and it is important that the utilities and BPU should be focused primarily on improving the design and execution of the programs, the *a priori* savings estimates for the next planning cycle, and targets for the program as it matures.

These guidelines are a set of study requirements to help focus the state and utility IPEs on how these evaluations of newer programs should be designed and implemented. Following this guidance:

- Utilities and the State and their IPEs 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. Utilities 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. IPEs are also encouraged to recommend specific scope enhancements that reflect priority issues for the program, or otherwise tailor the evaluation to make it most useful to improving performance of the program going forward.

These guidelines are not intended to limit the scope of the evaluation; they represent minimum expectations related to the impact and process elements for evaluation studies of new or transitioning state- or utility-run programs. These guidelines – and particularly the basic-rigor elements (sample sizes and counts / "desk review" elements) - are not intended for estimation of on-going programs, are not intended to meet PJM requirements, nor do they apply for evaluation of studies of impact parameters that do not vary with programs (e.g., operating hours studies, etc.), or certain other evaluations depending on start dates, topics, or impact/importance. Finally, there will be state- utility-run program evaluations that are suitable for higher-rigor evaluation. That will become apparent as we work with the IPEs on specific plans for specific programs, and those evaluations will be covered by separate Guidelines.

Note that utilities may, and are encouraged to, provide joint work plans for individual programs, as this would be expected to lead to evaluation study savings. However, the plans should include sampling plans that provide the recommended sampling precision for each utility. In addition, for the sake of evaluation economy, the guidelines include a recommendation for the study to collect of data to support NTG studies of the programs. Similarly, state studies are encouraged, where appropriate, to include sample sizes sufficient to provide utility-territory-specific results.

2.0 Impact Evaluations

Basic rigor impact evaluations are suitable for programs in start-up and transition phases; separate guidelines addressing Enhanced Rigor impact techniques are to be used for on-going programs, outside the transition phase. Impact evaluation efforts by IPEs for new and transitioning programs can appropriately be limited to basic levels of rigor because of the tendency of new programs to change rapidly as important lessons are learned, and because the programs are unlikely to have reached a stable configuration. Rigorous utility impact evaluation efforts in early stages thus have the potential to be quickly outdated by changes in program design and/or implementation.

For the purpose of basic rigor impact studies (sample sizes and level of investigation), the following summarizes the intent.

- Residential – All Measures: The IPEs will need to have access to data from the installer tracking database, because "count"-related tracked data on measures installed will be compared to the counts from survey responses, and the questionnaire needs to be tailored to the measures installed. The evaluations should also include subsampling as needed and feasible (by measure or by subgroup such as low income, geography, equity, or other factors).
- C&I – Verification-only analysis for Technical Reference Manual (TRM)[1] measures, deemed measures or for measures not included in the TRM. The verification efforts include the number of installations and the application of the proper deemed savings algorithm and parameters from the TRM. The Installation verification includes verification that the measure meets program requirements, was properly installed, and has the ability to produce savings. Sources for conformance for measures not included in the TRM include deemed criteria, or other applicable guidance as necessary. The analysis should include subsampling as needed and feasible (e.g., size of firm, size of project, type of building, Disadvantaged areas, etc.)
- C&I Verification-only analysis for custom or pay for performance measures. The verification focuses on proper measure installation and appropriate application of the approved audit and/or energy analysis provided with the application submission. Installation verification includes verification that the measure meets program requirements, was properly installed, and has the ability to produce savings. Verification of the custom energy savings analysis includes a desk review of the engineering methods used to estimate informed by the engineering specifications of the installed measure(s) and updated operating parameters based on actual installed equipment conditions. Where operating conditions can be reasonably obtained through customer interviews and do not need to be directly measured onsite, that is allowed. Where they cannot, evaluators should either identify a viable verification method or make a case for why the factor is not important and identify the range of error associated with omitting the verification of the item. No subsampling is likely appropriate for this group.

2.1 Impact Study Questions

The key study questions for the gross impact evaluation are:

- What are the differences between tracking-reported installations of specific measures and customer-verified measure counts installed? What are the percentage differences, by measure or (typical) measure groups?
- Using the TRM steps and values, or other available applicable guidance as necessary, what are the revised savings estimates, overall, and by measure or measure group? For custom C&I programs, was there proper application of the savings algorithms for low-impact custom measures using site data related to equipment characteristics that do not need to be measured on-site? What are the percentage savings differences overall, and by measure? What are the implications for realization rates?
- Are there elements of the computation that are not industry-standard? Are there TRM values or algorithms that should be updated? For commercial retrofit program, does the review of

[1] Consistent with Board approved methodologies adopted in the 12/2/2020 NJ Board Order, IN THE MATTER OF NEW JERSEY'S CLEAN ENERGY PROGRAM - FISCAL YEAR 2021 PROTOCOLS TO MEASURE RESOURCE SAVINGS, Docket No. QO20090584.

custom calculations identify issues or updates? Is the equipment still in place and operable? Has it been removed and why? Are there issues with the equipment or installations or training that should be addressed? What does this mean for the savings computations?

- (In the second evaluation) Were there improvements in the match between reported and verified values, were databases improved, or did highlighted issues or algorithm documentation or recommendations from the first evaluation lead to improvements?
- What are the primary drivers of the realization rates?

2.2 Sampling – Impact Evaluation

As mentioned above, these evaluations are intended to provide feedback during the early-program learning period. Therefore, the SWE recommends quarterly sampling and data collection. In addition, 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, 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.[2] For State-run programs, the evaluators will work with SWE to determine if the appropriate sampling is at the Statewide level, of if stratification to allow reporting out by utility territory is appropriate.

The evaluator should gather data from a statistically representative set of sample points (participants, etc.) for each measure of interest (e.g., those representing 5% or more of the savings, and/or at least 2 measures beyond lighting, whichever is more), comparing to the tracking data counts, and reviewing the TRM and calculation algorithms. Samples should be drawn randomly from each quarter's tracking data from the program, and data collection *and high-level analysis* should be conducted *quarterly* to support quick-turnaround correction of any key issues identified.

There must be a more detailed effort than verification counts for custom C&I, since by definition there is no applicable TRM. Here, desk reviews that incorporate interviews with samples of end users and suppliers / contractors must be conducted to provide appropriate verification of these savings. Interviews with the person(s) responsible for the calculations are recommended.

2.3 Additional Analyses Beyond Savings, ISR, RR

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

[2] If the population is too small for these requirements, a solution should be discussed with the SWE.

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.

3.0 Process Evaluation

The process evaluations[3] should be in-depth and focused efforts to ensure that the program problems have been uncovered and addressed prior to the next Triennium. The process evaluations should also support the impact evaluation effort to ensure that program tracking is in place and sufficient to support in-depth impact evaluation in the next evaluation or next Triennium.

3.1 Process Evaluation Study Questions

The study questions vary by whether the program is a new vs. a transitioning program.

3.1.1 Programs Transitioning in Management, Design or Delivery

The primary study questions for the first process evaluation of transitioning programs concern challenges and successes of the transition in management or major redesign of the program:

- Document what changes occurred in the program implementation and what stayed the same when the utility or state program transitioned in management, delivery or design.
- Document participation rate, closing rate, project completion rate, number of participants, and partial participants and, if possible, compare to previous implementation period.
- For partial participants (dropouts) identify reasons for failure to complete, reasons wanted to participate, challenges of participation, response to recommended measures (if recommended).
- Satisfaction with all key steps and elements of the program process by end users, reasons for participation, challenges to participation, decision-making, reasons for adoption or rejection of recommended measures, and suggestions to address challenges / barriers.
- Satisfaction with the back-office processes by the implementation team; cycle time findings for back-office processes.
- Satisfaction with all key steps and elements of the program processes by market actors involved in program delivery. For market actors involved in the previous program period, request assessment of any differences, their reasons for being in the program, challenges to participating in the program, access to products, reasons for recommending services and products, comparison of experiences prior to and during program, and suggestions to address challenges / barriers.

[3] These process evaluations should be formative, conducted during the operations period, not summative as in following the program operation period.

- Document any difficulties with program related efficiency products from end user, trade ally, and implementation team perspective such as availability, quality of materials, installation, or quality of product, waiting times, etc. Differentiate COVID related causes if relevant.

The study questions for the second process evaluation of transitioning programs will follow-up on findings from the first process evaluation and assess ability to address recommendations and achieve mature program status.

3.1.2 New Programs

The study questions of the first process evaluation for new programs introduced by the utilities or state focus on the success of the rollout and fine-tuning of the implementation and delivery process.

- Document the program theory and logic (note this is to ensure the program intention is clear and what should happen because of the program activities – outputs and outcomes)
- Document expected participation rates, closing rates, completion rates, time to completion, and dropout rates for participants and subgroups, as well as expected participation and engagement rates for trade allies.
- Document the status of program tracking databases and program participation forms and collateral; review collateral and marketing materials for effectiveness.
- Document status of back-office processes and marketing activities supporting the program, assess ability to fully track cycle time to manage the process.
- Document the end user experience with all key steps / elements of the program; reasons for participation, challenges with participation, reasons for adopting or rejecting recommended measures.
- Document the trade ally experience of the program, their reasons for being in the program, challenges to participating in the program, access to products, reasons for recommending services and products, comparison of experiences prior to and during program.
- Recommend how to improve the program processes to achieve the intended goals.

The study questions for the second process evaluation of new programs will follow-up on findings from the first process evaluation and assess ability to address recommendations and what additional research will be useful to ensure program is fully functioning in the following Triennium.

3.2 Sampling and Data Collection – Process Evaluation

The sampling and data collection approach varies for the two types of programs (new vs. transitioning). However, in both cases, SWE recommends that the samples should be drawn quarterly, and high-level results reviewed quarterly, to provide feedback to allow problems to be addressed promptly. If participation is too low to allow this frequency, other frequencies can be proposed. At the end of the program year, the evaluations should provide at least +/- 10% at 90% confidence overall, and 90% +/- 15% for each of the specific targeted subgroups/strata. These criteria are at the program level for each utility. For State-run programs, the evaluators will work with SWE to determine if the appropriate sampling is at the Statewide level, of if stratification to allow reporting out by utility territory is appropriate. 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.

Transitioning programs are expected to have large populations of participants. To address the study questions the evaluator should interview program staff and implementation staff, and survey samples of participants, partial participants and trade allies who support the program.

Survey and interview guides must use multiple survey items 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.

New programs will likely have smaller populations of participation. Therefore, the evaluator should conduct interviews with program staff, implementation staff, small samples of early participants, partial participants and trade allies who support the program in the first process evaluation. Surveys should be implemented as soon as participation rates allow, so process evaluation results can provide feedback to the program as soon as possible.

Interview guides should seek to uncover solutions to any challenges faced by the program and 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.

3.3 NTG Support: Sampling and Data Collection in Support of Statewide Net-To-Gross (NTG) Study

Economies can potentially be achieved if the IPEs augment the participant survey data collection work being conducted for the process (or impact) evaluation described above with a series of questions to support the NTG analysis. The goal is to provide NTG information at the utility territory level, and for whichever utility evaluation timings are suitable, collecting data as part of the Process evaluations will save evaluation money and avoid oversampling customers. The NTG evaluation work is expected to be a statewide study, with results reported at the utility level, using uniform questions and analytical methods. The NTG methodology is based on industry-best practice, relying largely on Massachusetts and other established methodologies, with some adjustments. For process (or impact) studies that are able to incorporate this data collection, the data collection of interest includes the following topics:

- Free ridership, identifying prior intention, with details related to timing, efficiency level, quantity, and influence factors.
- Inside spillover questions, which include a screener, and follow-up questions related to measures, measure efficiency, influence, actions in the absence of program participation, and some consistency checks.

The data, anonymized, is to be provided to the statewide NTG contractor for uniform analysis. The specific, uniform question batteries, which necessarily vary by program type, will be available from the

NTG guidelines. That study will be assembling data and collecting additional data, as needed, to support the NTG estimation. The NTG guidelines should be closely followed to ensure consistency.

3.4 Data Collection Instrument Format

The SWE recommends that data collection instruments 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.

4.0 Evaluability Assessment

Every basic rigor 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 enhanced 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 software or 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 basic rigor evaluation conducted on the programs.

5.0 Process / Impact Analysis Methods, Findings, Context, 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 jurisdictions. Therefore, each process and impact evaluation are 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 basic rigor results should include:

- Tables of verified in-service rates (with confidence intervals and sample sizes) at the measure level,
- Tables of reported savings and verified savings from the calculations performed, by measure, using all elements specified in the TRM[4],
- Tables of realization rates by measure, and
- Other information gathered in the study that provides performance results by measure, measure group, or program-wide.

6.0 Reporting

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

- A 1–2-page abstract including a 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).[5] 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

[4] Including heat / interaction effects, etc.

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

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.[6]
- 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 First Triennium and be available upon request (and without charge) to the BPU and their consultants.

6.1 Report Timing

The evaluation type and timing are addressed at the beginning of these guidelines. Obviously, evaluation results should be as current as possible. Given the relatively low-level of data collection and analytical requirements, basic rigor impact evaluations and process evaluations are likely to be able to be conducted on current program year participants, with timely results. For enhanced rigor evaluations, it is more likely that impact evaluations will be conducted on an earlier PY, but process evaluations should be conducted on the most recent PY or two of participants.

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

6.2 Style / communication:

- 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.

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

- 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.

7.0 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 (called Table 2)
- A data collection plan, including the data collection method for each group, and a table that identifies the key topics to be included for each survey / interview group (described below) (Called Table 3)

- 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 about how the measure counts will be verified, and the steps anticipated to assure as collection of as accurate data as possible.
- Detail about 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 staffing plan (hours by 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 (Table 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 (Tables 2 and 3 in the scope),
- 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 (Called Table 1 in scopes)

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	MA, MD, CT, CA		
IMPACT EVALUATION			
Process & impact together?	Yes	No	
Program Year	2		
Study Start / end date	7/23-12/23		
Solo or with other utilities (list)	Across all		
Rigor level	Enhanced		
# program participants expected	600		

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
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	MA, MD, CT, CA		

8.0 References

Questions: Contact Jane Peters (JaneStrommePeters@outlook.com), Lisa Skumatz (skumatz@serainc.com), or 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.