New Jersey Evaluation Guidelines: Process and Impact Evaluations for Behavioral Programs May 2023

Guidelines for Clean Energy Program Evaluations

Sector: Residential Behavioral Programs using home energy feedback reports

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
Revised from April 12, 2023

Contents

Abstract

This document focuses on guidance regarding program evaluation efforts by utility and state Independent Program Evaluators (IPEs) for residential behavioral programs delivered using home energy feedback reports. 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.

If a question arises about what is intended to be covered by these guidelines, the SWE will clarify as needed for each instance.

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 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

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. 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.

2.0 Objectives

The purpose of a behavioral program evaluation is to estimate savings attributable to the program, and to provide recommendations for improving the delivery and effectiveness of the program. The study[1] should also differences in savings between treatment groups. Treatment groups[2]are studied because savings tend not to be uniform across income and usage-level groups. New Jersey will benefit from understanding the differences in actions, barriers, and savings between these groups. In addition, the influence of behavioral programs on other programs in the portfolio are important to provide appropriate credit and avoid double-counting of savings effects. The study should be designed to produce results that can help identify possible gaps in messaging between groups and identify most effective messaging to improve impacts and reduce messaging that focuses on approaches that are barriers for some groups, and help clarify the savings effects attributable to the program and ways to reinforce and enhance the underlying induced behavioral changes.

Results are desired in two main areas: impact-related and process-related.

Impact-related results are estimated via billing analysis using control and treatment groups; and these guidelines requests additional results by subgroups.[3] Results should include:

- Estimated savings impacts due to the program, where possible, delineated by pre-usage level, and possibly by income classifications.
- Where possible, (where cohorts have stopped receiving feedback reports) persistence of savings;
- Estimates of cost per savings,[4] assessing the ratio of program expenditures to savings demonstrated, potentially by the subgroups.
- In later years, examination of differences in savings from homes receiving reports for 1 year vs. 2 years, etc.

Process-related, using phone or web surveys, plus focus groups, with results for the subgroups, should be produced, including:

- Report usefulness to the households, including the role of the report, comparisons to neighbors, barriers to use, who read the report,
- engagement with any electronic portals,
- awareness / recall of the program and its components, learnings about other programs, and implementation of tips,

[1] "The study" in this document always refers to the evaluation study.
[2] Referred to as subgroups in this document, post-identified after the RCT and overall treatment / control are selected.
[3] It is recognized that the precision levels associated with conducting these subgroup analyses will be lower and that it "breaks " the RCT. Checks for comparability for each subgroup between the treatment and control groups are desirable to provide additional confidence in the results. This is acceptable for these important analyses.
[4] With necessary or appropriate caveats; specifics or alternatives can be discussed with the SWE during the scoping phase.

- Information on the attributable actions taken at the household level and uplift to other programs[5]
- and other topics.

Activities, methods, and outcomes are discussed below.

Evaluators may suggest alternative approaches to achieve the desired outcomes and results during the scoping process and discuss the rationales with the SWE for consideration.

3.0 Methods

3.1 Sampling and Control Groups

Utilities typically implement behavioral programs with home energy feedback reports as randomized control trial (RCT) or randomized encouragement design (RED) experiments. In each case customers in a homogeneous sample frame are randomly assigned to treatment and control groups. The evaluation sample is the original implementation sample frame including any drop outs, outliers, or other customers removed by the implementer.

The study must be designed to include subsamples by usage level; if possible, SWE would also be interested in subgroups defined by income group. Usage levels are available from billing records; income group can be approximated by use of census block group level. The study should include the ability to represent statistically valid results for:

- Representative group across all usage levels;
- Highest 1/3 of usage levels
- Model 1/3 of usage levels, and
- Lowest 1/3 of usage levels.

If the sample is not pre-designed in this manner, then post-stratification should be used to identify differences at least by pre-usage levels within the sample.

Control Groups

Suitable control groups are required to conduct the evaluation of these programs. If the control group is already selected (for example, by the implementer[6]) and fixed, the suitability of the control group must be checked by the evaluators. This preferably involves obtaining the files related to creation of the RCT scheme and reviewing the selection procedures. Whether or not these files are available, it is also necessary to conduct statistical work to checks for the match between treatment and control groups. This includes comparing the two groups on independent data including usage levels, geographic

[5] Two levels of this feedback are obtainable. A participant survey will provide useful information on actions households attribute to the program. A defensible analysis of the manner in which the program causes savings would require a substantial non-participant survey. This is desirable, but not required in these guidelines.
[6] These guidelines direct the actions of the evaluators, not the implementers.

distribution, and other factors (e.g., weather zone, etc.). If the groups are not well-matched, the savings results will not be reliable. In this case, the evaluator should discuss options with the SWE.[7]

If the evaluators are supplied with data and procedures related to the creation of the control group, a number of items about the construction and features should reviewed, including:

- Construction using RCT or RED methods.[8]
- Size should be consistent with PA SWE Framework Chapter 6[9]
- Statistically similar to the treatment group(s) in demographics and usage, using a combination of randomization, propensity scoring, and other factors (equivalence tests).[10]
- Non-participants / no messages or feedback reports
- Preferably, a year of "pre" billing data
- The control group should be updated as needed, using the same principles.

Again, the evaluators are required to check the comparability of the control and treatment groups.

3.2 Process Evaluation Methods

The evaluators should use three methods to inform the study objectives:

1. Review of program materials and tracking data
2. Telephone or web-surveys with participants (treatment), and preferably, also with non-participants (control) from the different subsamples to more rigorously assess impacts of reports on customer behavior
3. Focus groups (in-person or virtual) with participants from the different subsamples

Review of materials and tracking data: A basic approach for process evaluations is review of program materials and collateral, as well as an analysis of any program tracking data. Assessments of clarity, consistency, effectiveness and other aspects of the material are expected. These activities are expected in the evaluation of this program.

Telephone or Web Survey: The sample sizes provided assume that the program is administered to thousands of households in the territory. Given the large samples sizes for these programs, the overall sample should provide precision / confidence at the level of 5/95 at the program-level, with goals of 10/90, if possible, for each subgroup. The survey instrument should be constructed after a detailed discussion of key issues is held; SWE should review the instrument.

The survey should focus on the following issues:

[7] It is desirable that the utilities require all control / treatment group creation files as part of implementation contracts. This has been a consistent problem in other states, and undermines the reliability of the evaluation.
[8] Quasi-experimental methods are sometimes used, but are less desirable.
[9] Sample size is generally greater than 10,000. Customers, smaller if a bootstrap analysis of the combined control/treatment frame shows that it will meet a 0.5% absolute precision requirement. See *Evaluation Framework for Pennsylvania Act 129 Phase IV Energy Efficiency and Conservation Programs, July 2021*, https://www.puc.pa.gov/media/1584/swe-phaseiv_evaluation_framework071621.pdf.
[10] Construction of treatment / control group subgroups by usage level and other factors with sufficient samples sizes in each subgroup would be desirable.

- **Engagement with the program,** including an assessment of whether treatment households were aware of the program and read the feedback documents, which household members were engaged, reported usefulness of the information, messaging recall / engagement, perceived relevance of the energy-saving tips, and use of on-line resources provided by the program. A key element should be changes over time, and whether readers become tired of the reports and their effect changes over time.
- **Engagement with the reports:** including respondents' reactions to the feedback reports, changes in readership over time[11], whether they pursued additional information; and other open-ended questions.
- **Barriers to engaging in the program**, including household situation (owner / renter), income, ease of understanding the reports, applicability of recommended tips, access to internet / digital aspects of the program, time and other factors.
- **Evidence of behavioral change** including self-reported information on respondent behavior related to energy use providing context for electricity and/or gas savings analyses. Questions may include whether treatment households take steps to reduce their energy use as a result of the reports, whether they follow-through on energy-saving ideas presented in the reports, whether household members discuss energy conservation and the ways they can reduce the household's energy use, whether it led them to participate in other programs and which program(s)[12], and other information.
- **Satisfaction with the program,** including respondent satisfaction with the reports as well as self-reported attribution as to whether the program had helped reduce their energy-use. A series of open-ended questions should be included to point to ways the respondents believe the program could improve.
- **Demographics** should be included in the survey, so the analyses can examine statistical differences between the subgroups. Previous studies have identified differences in the following (which should include: Households with children; Households with residents 65+; Number of people living in household year-round; Own or rent home; Type of home; Year home built; number of bedrooms in home; Education level; household income.

Survey methods may include advance letters or postcards, or incentives; they should include multiple attempts (at least 3) before the selected household is abandoned. The study must report final dispositions including response and completion rates by group; results should be reported with confidence intervals.

Focus Groups: Screening for the focus groups should include asking if household members were aware they were receiving feedback reports, and only individuals who indicate that they were aware of the reports should be included in the focus groups. Distribution of attendees by race, gender, age, income, own/rent, and other categories are recommended. The focus group facilitators can determine length (90 minutes is recommended), and whether and which observers may attend. The primary objectives of the focus group discussions are to:

[11] Based on current responses for initial evaluations, augmented by repeated metrics over time as part of later evaluations of the program.

[12] And preferably, also measures installed and incentives taken

- Identify attendees' evaluations of the behavioral program and the home energy feedback reports
- Usefulness of the report information for their household, and which household members read the reports
- Levels of readership and engagement with the reports, and whether that has changed over time / tiring of reports
- Ideas for changes in the program that could increase engagement and satisfaction
- Suggestions for changes in presentation or messaging
- Drill-down on barriers

The focus group script should be constructed after a detailed discussion of key issues is held; SWE should review the script. A report that is action-oriented is desired. While it cannot be statistically-representative, it should provide key findings and identify any consensus items. Recommendations should be provided. The purpose of the focus groups differs from the survey in the access to drilling down on issues, rather than receiving only high-level answers. That drill-down information should be a part of the study.

3.3 Impact / Billing Analysis Methods

The evaluator's billing analysis should focus on the electricity and/or gas savings induced by the program. As mentioned, the results should be presented overall, as well as by usage (and potentially income) subgroups.

The impact evaluation sample is the original implementation sample frame including any dropouts or other customers removed by the implementer.

There are several steps related to data preparation that are needed prior to performing the statistical analyses.

Billing Analysis Data Preparation. The billing analysis relies on multiple pieces of data. The discussion below addresses monthly data, but note that the billing reads may not be strictly monthly, and that is acceptable. Data needs include:

- The utility, for flags for households that opted out of the program, flags for program participation (by year and type if possible), rate codes for all-electric rate paying households, and flags for and dates of service disconnection.
- The behavioral program administrators, for billing data files matched to participation (and control) account numbers. This includes monthly electricity and/or gas usage as well as meter read dates, flags for estimated meter reads, and dates reports were sent for each household.
- For some modeling approaches, the National Climate Data Center (NCDC) website, for weather data for relevant regional stations within or near New Jersey. The service account zip codes need to be matched to the nearest weather station. For each region, the evaluators need to

calculate average monthly temperature, total monthly heating degree days, and total monthly cooling degree days from daily data available from the NCDC website for the study period.[13]

Data cleaning of other types are also needed. Households may need to be removed for lacking a match to billing data, service disconnections, lacking unique billing accounts, and other reasons. The final database should include household characteristics, rate code, monthly billing data, and monthly regional weather data with sufficient sample sizes to conduct the analysis.

The study should provide analytical results, as well as overall and monthly electricity and/or gas use for the households in each treatment group across the study period, pre-program and post-program, with information indicating groups that have values that are statistically significantly different at the 90% confidence level.

Overall Program Savings Estimation Procedure

Evaluators should address the following questions prior to determining ex post savings:

- Was the original sample size large enough to detect the savings target?
    - Sample design should result in 0.5% absolute precision at 95% confidence at the program level. Precision for subgroups may be lower. Depending on the targeted per customer savings sample sizes are typically 10,000 or greater.
- Are the treatment and control groups equivalent?
    - Average pre-treatment energy use and distribution should be similar for both control and treatment groups.
    - If additional characteristics, such as floor area, household income, weather[14] or weather against heating fuel, or other factors, are available they should also be similar for both groups.

Modeling Approaches:

The evaluators should use customer energy bills to determine whether the program has successfully resulted in behavior change and long-term reduction of energy use. The team should estimate energy savings and the persistence of savings through the use of billing analysis.

There are multiple specific modeling approaches available for these programs, and the selected approach should be identified in the scope and discussed with SWE. Some are weather independent; others incorporate weather data, for instance. Examples of several acceptable modeling approaches are outlined below.

1. Robust OLS: Standard billing analysis methods rely on ordinary least squares (OLS) robust regression, which is resistant to any imbalances in pre-program use between treatment and control groups and also to data point outliers; thus, OLS ensures that the method does not over-estimate or underestimate treatment effects. Robust OLS can also adequately handle various

[13] This weather normalization step may be omitted if the control group / treatment group match on this fact can be checked and confirmed by the evaluation team.

[14] evaluators should consider weather normalizing when the weather seen by the sample frame is not homogeneous.

missing data effects.[15] To use the billing analysis to understand energy savings for specific time periods and sub-groups, the treatment and control groups can be divided into various subgroups by restricting the data by time period or characteristic of interest.

2. Lagged Dependent Variable: After the check for treatment/control imbalance using energy usage and other available data, and the verification that pre-period energy usage is statistically equivalent between the treatment and control groups, the modeling is conducted. Savings are estimated using a Lagged Dependent Variable (LDV) model[16] to produce unbiased estimates of program savings, and these estimates can be broken down by previous treatment status, fuel type, pre-period usage and other factors. The LDV model combines both cross-sectional and time series data in a panel dataset. The modeling approach uses only the post-program data, with lagged energy use for the same calendar month of the pre-program period acting as a control for any small systematic differences between the participant and control customers. The model derives an estimate of average daily energy savings attributed to the program. Total energy savings is calculated as the product of the average daily energy savings and the total number of participant-days in the analysis. For each fuel type, differences in program outcomes for usage groups can be examined, and the evaluators can test versions of the model that include additional covariates (e.g., prior participation status).

Other models may also be considered. Evaluators should identify the modeling approach they plan to use within the scope.

For any of the estimation methods used, subgroup analysis is of interest. As mentioned, pre-usage categories are of particular interest, as previous work indicates very different savings levels based on this variable, which affects conclusions regarding program expansions. time periods of interest may include summer or winter months, or households paying the all-electric rate or being high / medium / low users. For the persistence analysis (if relevant) variations in the frequency of how often reports are sent out can make a difference and be an important analytical variable.[17] An outlier analysis or other special treatments or analytical techniques may also be appropriate.[18]

3.4 Additional Analyses beyond Savings

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

[15] For example, inadequate post and pre-treatment electricity and/or gas use information as well as households lacking treatment/control assignments that may not be evenly distributed between the treatment and control group households. This creates an imbalance in the dataset, and robust OLS addresses such imbalances.

[16] Savings estimates produced by the LDV model tend to be the most accurate and precisely estimated compared to alternative model specifications. Assuming the RCT is well-balanced with respect to the drivers of energy use, the LDV and alternative models should produce similar program savings estimates, increasing confidence in evaluated savings.

[17] historically, for example, some utilities have had groups with monthly vs. quarterly reports

[18] It is recognized that the precision levels associated with conducting these subgroup analyses will be lower and that it "breaks " the RCT. This is acceptable for these desired analyses of subgroups.

Persistence Analysis: A persistence analysis is desired if possible; this has been valuable in other states. This can provide defensible program lifetime information for the TRM. In addition, analysis of the persistence of savings after households stop receiving reports (if any cohorts are, or can be, cycled out of the program)[19] can provide data on the optimal operation of the program (cycling) for greater cost-effectiveness and refreshing of the impacts.

Uplift Analysis: An analysis of the impact that the behavioral program and tips has had on other programs is expected. The analysis will use survey responses to questions about in measure installations, participation in other programs and associated incentives. Reliable information is obtained by examination of differences between responses from participants and non-participants; less reliable and indicative information can be obtained from participant-only surveys with self-reports on the influence of the behavioral program on these follow-up actions. The former method is preferred.

NEB/NEI analysis: Non Energy Benefits (NEB) / Non-Energy Impact (NEI) analyses have rarely been conducted for these programs, due to the mostly-indirect behavioral and measure installations. This program may benefit from this estimation, and the relevant questions may be incorporated into the process evaluation survey. To help aid this and the uplift work, the process and/or impact surveys should collect information on the specific behaviors taken and measures installed as a result of the feedback and tips from this program.[20] The evaluators and SWE should discuss the approaches to incorporate a NEI analysis, and a discussion included in the evaluation scope.

3.5 Outputs of Interest from the Study

Process outputs should include recommendations on design of reports, tip tailoring to groups, analysis of barriers for certain groups and retailoring of the program to make the program more effective and meaningful, and other outputs. Impact-related outputs can include:

- Program-attributable energy savings overall and by subgroups, with inferences regarding direction of overall savings as different subgroups are targeted or added to the program.
- Program overlap and indications of uplift-type effects the program may be able to claim via induced participation in other programs. Serious attribution will take a drilled-down version of the survey questions regarding actions taken.
- Persistence work can help inform appropriate lifetimes for the program for the TRM.
- Dividing program savings by costs overall and within group can provide an indication of the relative cost-effectiveness of the program for specific groups. If program records vary in frequency, variations in cost-effectiveness may also be analyzed and used for program refinements, including possible "cycling" (periodically taking customer cohorts off the program to refresh the attention to the messaging), or other changes.

[19] The SWE is interested in persistence analysis as the opportunity arises. These results will inform updates to the TRM for EULs when treatment is discontinued.

[20] Whether the participants installed received incentives from other programs should also be collected.

4.0 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 a Behavioral Program 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 of data (not individual variables) that were verified, and that the IPE confirms that the data to support Behavioral Program Process and Impact evaluations 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 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 program.

5.0 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 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 where relevant:

- 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: savings results, and 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 wave or subgroup
- 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, messaging, targeting for maximum impact, and recommendations for improvement to 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 behavioral results should include:

- Tables of gross and/or net savings and realization rates by wave and subgroup; [21]
- Persistence, uplift, and other information gathered in the study that reflect program performance results.

[21] Including all appropriate adjustment factors in the TRM

6.0 Reporting

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

6.1 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).[22] 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 for benchmarking and 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.

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

- It is required that all data purchased for the project becomes the property of or accessible to all other NJ evaluations.[23]
- 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.

6.2 Report Frequency

Obviously, evaluation results should be as current as possible. For the Behavioral Program, both impact and process evaluations can be conducted on the most recent program year (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.

6.3 Report Timing, Data, and Data Collection

The goal of the impact evaluations is to provide the final reports in time for inclusion in the Evaluation Use Memo, which includes all reports completed by December 1 of each year. This is the cut-off date for NJ research study values to be included in the Updated TRM.

Special considerations for data issues include:

- Timing and schedules for data-driven impact or process evaluations might deviate from the expected schedule. For example, A heat pump impact study requiring twelve months of metering may 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.

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

6.4 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 are 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.
    - 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 and other 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 (Table 2) 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, and a table that identifies the key topics to be included for each survey / interview group (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 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 (Tables 2 and 3 of the scope)
- A discussion of risks and how they will be addressed,
- A list of utility and other data to be requested,
- A succinct discussion of each task and how the analysis will 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. Include approximate population sizes for the survey and interview groups.

Figure 1: Evaluation Studies Summary Table (Table 1 of the Scope)

EACH COLUMN is a separate study. *Abbreviation "N"=Number of observations*	Program, PY & Study Name (sample answers) Behavioral Program 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	Behavioral		
# program participants expected	30,000		
Program's expected share of portfolio savings	10% of portfolio, 50% residential		
Program documentation to be reviewed (tracking, outreach, web, etc.)			
Staff, method (~N)	IDIs/ ~5		
Participant method, (order of magnitude N or precision/confidence),	Web Survey, stratified by pre-usage, combined with Impact, 380 sample; ALSO focus group of participants		
Partial Participant method, (order of magnitude N or precision/confidence)	No partial participant survey, or should we say opt outs should be done –I'd say not		

EACH COLUMN is a separate study. *Abbreviation "N"=Number of observations*	Program, PY & Study Name (sample answers) Behavioral Program Impact & Process	Program, PY & Study name (example for a process-only study)	Next study / Study for Program 2
Non-Participant (order of magnitude N? or precision/confidence)	Yes, control group sample web survey size is 380 for uplift analysis.		
Vendor / contractor surveys (N/precision), specify group / groups	None		
Measure or end uses? (specify key ones)	N/A for behavior; info on behaviors & measures installed through participant process / impact surveys		
NTG survey included? How many "N"?	Not applicable		
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).*	No		
Other notes, items included…			
Date and PY for last process evaluation	6/22-12/22, PY1	None	
Rigor level for last previous evaluation	Behavior	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		
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		
Partial Participant (order of magnitude N or precision/confidence)	No		
Non-Participant (order of magnitude N or precision/confidence)	Yes, important – web survey of 380 from control group with process		
Vendor / contractor (N/precision), specify group / groups	No		
Measure or end uses? (specify key ones)	N/A		
In-service / verification planned? N, Method	N/A		

EACH COLUMN is a separate study. *Abbreviation "N"=Number of observations*	Program, PY & Study Name (sample answers) Behavioral Program Impact & Process	Program, PY & Study name (example for a process-only study)	Next study / Study for Program 2
Impact evaluation(s) method planned	Billing analysis OLS		
TRM generation applied	2022 Comprehensive		
NTG survey included? How many "N"?	Not applicable		
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)*	Large vs. smaller pre-use categories; income/disadvantaged areas		
Other notes, items included…			
Date and PY for last process evaluation	6/22-12/22, PY1		
Rigor level for the previous evaluation	Behavioral guidelines		
Was evaluability certified in last evaluation?	Yes		
Other evaluation type			
States/utilities for comparison (included in body of report)	MA, MD, CT, CA		

8.0 References

Questions: Contact Jane Peters (JaneStrommePeters@outlook.com) or Lisa Skumatz (skumatz@serainc.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
e. The Uniform Methods Project – https://www.nrel.gov/docs/fy21osti/77435.pdf

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