EVALUATION STRATEGIES FOR THE SITE-SPECIFIC SAVINGS PORTFOLIO

Submitted to

LAUREN GAGE
Bonneville Power Administration

Submitted by

SBW CONSULTING, INC.
2820 Northup Way, Suite 230
Bellevue, WA 98004

In association with

RIDGE AND ASSOCIATES
ENERGY & RESOURCE SOLUTIONS
BUILDINGMETRICS
JACOBSON ENERGY

January 19, 2017

sbw

ENERGY + WATER + EFFICIENCY

NOTE FROM BPA

BPA would like to clarify the use of this memo. This memo provides the background we need to further develop BPA's Integrated Evaluation, Measurement and Verification and Quality Control approach for custom projects.

BPA's integrated design approach aims to shorten the amount of time it requires to gain useful feedback to energy efficiency program designers as well as to reduce effort from measurement and verification, quality control, and evaluation. In addition, we expect the approach to enable evaluation staff to acquire better project baseline data, as well as to reduce the need to contact the project end users or BPA's utility customers for the needs of evaluation years after the project was complete.

In order to develop such an approach, BPA needed to: a) understand key components of evaluation, b) understand efforts and results of other energy efficiency organizations across the nation regarding "real time evaluation", and c) gain insights from the expert evaluator commentary on our initial ideas. This memo provided this background understanding to continue to design our Integrated Evaluation, Measurement and Verification and Quality Control approach for custom projects.

TABLE OF CONTENTS

NOTE FROM BPA II

1. BACKGROUND AND PURPOSE 1

2. DEFINITION OF EVALUATION 2

2.1. Relevant Portions of the RTF Guidelines 3
2.2. Summary of National Guidance 3

3. DEFINITION OF "REAL-TIME EVALUATION" 5

3.1. Opportunities for Real-Time Evaluation 6
3.2. Summary of National Experience 7
3.3. Benefits of Real-Time Evaluation 8
3.4. Risks of Real-Time Evaluation 8

4. BPA'S REAL-TIME EVALUATION OPTIONS 9

4.1. Possible Evaluation Designs 9
4.2. Role of Third Parties 11
4.3. Program Delivery Channels 12

5. M&V 2.0 IS A SEPARATE CONCEPT 12

A. ANNOTATED CITATIONS - DEFINITION OF EVALUATION 14

B. ANNOTATED CITATIONS - REAL-TIME EVALUATION 18

1. BACKGROUND AND PURPOSE

From June 2013 to December 2015, BPA conducted an impact evaluation of its site-specific savings portfolio. The portfolio contained all non-residential lighting and custom projects completed during FY 2012 and 2013 by both Option 1 and Option 2 utilities. An independent third-party evaluation contractor had lead responsibility for evaluation design, implementing the design, and reporting, however, BPA staff were substantially involved in providing guidance on research objectives and critiques of technical methods and report products. For all but one element of the portfolio, the evaluation largely validated the program savings claim. The evaluation took a long time to complete, consumed considerable time from BPA, utility and end user staff, and to some, it appeared to duplicate the quality control (QC) and Measurement and Verification (M&V) activities routinely carried out by BPA and utility program staff.

The recently completed portfolio evaluation was largely consistent in its design and methods with evaluations conducted by many electric and gas utilities throughout the US. The evaluation was conducted by an independent third party and managed at BPA by the evaluation lead. The evaluation design was post-claim, which means that the evaluation sample was not selected until after the program had claimed savings for a period of operation. In many jurisdictions, such evaluations have been routinely conducted, for a decade or more, by independent third-parties. In the case of BPA, there had never been an evaluation of this scope so the whole process was novel to many of the BPA and utility staff involved in the work.

As is common, the impact evaluation had two primary goals. The first was to verify the program's claimed savings. The second was to identify strategies for improving program performance, specifically, the program's ability to reliably estimate savings. For regulated investor-owned utilities, savings must be verified to satisfy regulatory requirements. BPA has a similar responsibility, under Northwest Power Act, to acquire reliable savings. In both cases, the second goal is important as it reduces the risks of unreliable future savings estimates. However, the long duration between completing any project and receiving evaluation results reduces the value of the evaluation findings and makes the evaluation process more difficult. Program procedures and staffing continuously evolve and the staff involved in a sampled project may no longer be part of the program staff by the time feedback comes from the post-claim evaluation.

The central question addressed by this paper is whether there are viable alternatives to post-claim evaluation designs that would allow BPA to meet one or both of its evaluation goals. Specifically, this paper explores the idea of "real-time" evaluation. Ideally, real-time evaluation would meet both evaluation goals while providing timely feedback to programs with reduced effort by programs and end-user staff.

We start by defining "evaluation" and grounding that definition in national practice. Next, we define "real-time" evaluation, describe to opportunities for real-time sampling, summarize the national experience with such evaluation designs and highlight the benefits and risks for these

designs. Finally, we discuss BPA's evaluation options, the role of third parties in implementing evaluations and constraints imposed by BPA program delivery channels. The Appendices provide two annotated bibliographies listing papers and reports relevant to the definition of evaluation and experience with real-time evaluation.

2. DEFINITION OF EVALUATION

There appears to be a national consensus that an evaluation is a process conducted by researchers that do not have any stake in the results of the evaluation. In particular, they do not personally benefit from the program's existence[1]. If the researcher is involved in the design, planning, operation or supervision of the program they may bring a bias, intentional or not, to an evaluation of the program's performance. Good evaluation practice requires minimizing the likelihood of the bias created by the influence of those responsible for operating the program. This is consistent with the basic description in Section 8.3.2 in the SEE Action Guide which states that the role of an (independent) third-party evaluator is to produce estimates of savings that all parties to the evaluation believe are based on valid, unbiased information that is sufficiently reliable to serve as the basis for informed decisions.

We believe there is an emerging consensus that for impact evaluations, "independent" means not working in the same organization as the program implementer. It also means that the evaluator has no financial stake in the evaluation results, which would create bias in favor of or opposed to the interests of the program implementer, program participants, or other stakeholders such as utility customers (consumers). A key issue is whether staff, working in an evaluation department, are sufficiently independent to conduct evaluations. With one exception, we find general consensus that staff in a separate department, which has no responsibility for program implementation, are sufficiently independent to conduct or manage evaluations. That one exception has been created by the California Public Utilities commission, which prohibits all staff working for the state's investor-owned utilities, regardless of their department, from conducting or managing impact evaluations.

It is difficult to achieve a consensus regarding the other evaluator characteristics such as affiliations, skills, and experience. There is a general consensus that an independent third party evaluator should possess all the necessary skills to measure/estimate specified parameters at a level of reliability consistent with the evaluation's budget. This is consistent with the *AEA Guiding Principle* that evaluators are expected to conduct systematic, data-based inquiries about whatever is being evaluated and provide competent performance to stakeholders.

The national consensus is less clear on what skills an evaluator needs to conduct a reliable evaluation. Engineering, economic and statistical modelling, sampling, interviewing, physical inspection and measurement are all relevant. However, the specific skills needed to evaluate an upstream lighting program are substantially different that those needed to evaluate custom

[1] Of course, this is an ideal. Even third-party researchers are compensated for conducting evaluations. If the programs did not exist, they would not be hired to evaluate the program, so any finding that leads to the cessation of the program is against their personal interests. But the intent of the national consensus is clear.

industrial process improvements. The RTF Guidelines provide a listing of relevant skills that is useful, but still will not substitute for defining the specific requirements of each evaluation.

2.1. Relevant Portions of the RTF Guidelines

The RTF Guidelines do not provide a formal definition of "evaluation," but portions of the opening paragraphs of section 5 in the *Guidelines for the Estimation of RTF Savings* provide a partial definition:

> Program impact evaluations estimate savings from a period of program operation. … Impact evaluations should be designed to achieve reliable estimates of savings while accommodating the special requirements of the program's delivery methods, target markets, efficiency measures, operating agency, and regulatory environment.

Although, "independent third parties" are listed as an audience for the Guidelines, section 5.2 states that evaluations are conducted by a "team of professionals." That section goes further to describe the required skills of that team. The team should be able to successfully perform the following tasks:

- Select representative and efficiently designed samples, i.e., maximizing precision for a given sample size.
- Collect and prepare analysis-ready site-specific data, e.g., surveys, inspection, measurement and billing data.
- Estimate savings using a variety of engineering and statistical techniques for sampled measures.
- Extrapolate sample findings to the study populations and quantify the uncertainty in this extrapolation.
- Prepare transparent and clearly written report describing the study methodology and findings.

2.2. Summary of National Guidance

The *Energy Efficiency Program Impact Evaluation Guide* (SEE Action Guide) (pp. xvi and 8-5) has proposed the most useful definition of Evaluated Savings:

> Values reported by an independent, third-party[2] evaluator after the efficiency activities and impact evaluation have been completed. The designations of "independent" and "third-party" are determined by those entities involved in the use of the evaluations and thus may include evaluators retained by the program administrator or a regulator, for example.

[2] Note that the term "independent, third party" as it is currently used in our field suggests that the two terms are essentially synonymous or that the two concepts are inseparable.

On p. 3-2, they provide a rather circular definition of an independent, third-party evaluator as "As an entity that conducts evaluations and is designated to be independent of the implementer and administrator."

On p. 8-9, the Guide states:

> . . . there is no formal definition of independent or third-party evaluator, as well as there are no well well-established precedents as to who hires the entity(ies) that provides the evaluated savings reports. The hiring entity could be the regulator or the administrator, or perhaps some other entity.

However, they go on to state:

> . . . in general practice, "independent third party" is thought to mean that the evaluator has no financial stake in the evaluation results (e.g., magnitude of savings) and that its organization, its contracts, and its business relationships do not create bias in favor of or opposed to the interests of the administrator, implementers, program participants, or other stakeholders such as utility customers (consumers). However, different states' regulatory bodies have taken different approaches to (1) defining the requirements for evaluators who are asked to review the claimed savings and prepare evaluated savings reports, and (2) who hires that evaluator.

Also, note that establishing independence is not as straightforward as it might appear. For example, the CPUC Energy Division (ED) has determined that it is impossible for any Investor Owned Utility (IOU) to manage an independent, third-party impact evaluation even with rigorous ED oversight. In the past, the IOU EM&V departments were considered sufficiently removed from the design and implementation of the IOU energy efficiency programs that they were considered effectively to be "third-party" as long as there was rigorous oversight of the IOU-led evaluations. That the definition can change over time even within a given jurisdiction is consistent with the Guide's statement above (pp. 8-9) that there is no formal definition; it all depends on the context.

This definition in the SEE Action Guide was considered sufficiently detailed to be adopted by the Uniform Methods Project (see Chapter 1: Introduction). However, the UMP doesn't provide the additional detail provided by the SEE Action Guide that "independent third party" is thought to mean that the evaluator has no financial stake in the evaluation results."

The *Impact Evaluation Framework for Technology Deployment Programs* notes:

> In implementing evaluations, program managers need to maintain an "arms- length" relationship between evaluators and themselves. This creates an extra burden on the part of programs to create transparent and defensible evaluation processes and conduct quality evaluations using independent evaluators. (p. 1-1)

They go on to say that to evaluate a technology deployment program, an organization should:

> Engage a qualified and independent contractor to conduct the evaluation. The evaluation contractor (or the evaluation contractor's firm) should not be involved in program planning or program implementation except when invited to provide insight or to monitor planning or implementation activities. The contractor should keep the sponsor informed when the

> contractor is asked to interact with program implementers to avoid the appearance of conflicts of interest. (p. 3-25)

Both the American Evaluation Association and the *Program Evaluation Standards: A Guide for Evaluators and Evaluation Users* provide much a more detailed and nuanced assessment of what it means to be an independent evaluator.

Other guidance documents were less helpful. The *California Evaluation Framework* only stipulates:

> . . . that program evaluations will be conducted by firms, organizations, or groups that are independent of the implementation administrator or contractor and that the evaluation teams will maintain an arm's-length relationship with implementation administrators and contractors in order to help assure objective and reliable evaluation efforts. (p. 21)

The *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*, while mentioning "independent evaluator" 38 times, never provides a formal definition.

The RTF's *Guidelines for the Estimation of Energy Savings* mentions independent third party but never defines it.

With respect to the other characteristics, all of the documents lay out the statistical, engineering and social science skills in varying levels of detail required to conduct an impact evaluation and most refer to IPMVP or the Federal Energy Management Program (FEMP) M&V Guidelines as the foundation for on-site M&V. For other skills, such as sampling, most of the resources, e.g., California Protocols (see p. 45 *Guidance on Skills Required to Conduct Impact Evaluations)* and the SEE Action Guide, are more general and provide general sampling definitions, principles and references that should be followed while others are more detailed (*California Evaluation Framework* and the *Sample Design Cross-Cutting Protocols* in the *Uniform Methods Project*) and provide step-by-step procedures and formulas.

3. Definition of "Real-Time Evaluation"

The term "real-time evaluation" is not well defined. In an industrial context, it might refer to continuous sampling and measurement of product quality, e.g., scanning welds for faults or testing the response of electronic components, at points in a production line. Such evaluations may involve high sampling rates, measurements with known reliability, and little opportunity for measurement bias. These measurements could provide almost immediate feedback to those managing the process. This type of evaluation has a single goal, which is to improve the process, i.e., make it more likely that future production will meet the manufacturer's product standards. It is analogous to evaluating a sample of proposed efficiency projects before they are implemented to determine whether the model of energy savings is reliable and likely to result in the estimated savings.

The manufacturer has some concern about historical levels of product quality[3]. But, the focus of real-time evaluation on the factory floor is solely on future product quality and is not designed to draw conclusions about last year's production. In particular, there is no burden to statistically estimate the quality of historic production, as must be done to meet the first goal of evaluation discussed above - to verify savings from a prior period of program operation.

To draw conclusions about the quality of production for a prior period, the evaluation sample must be statistically representative of the entire production during that period. This is in addition to the requirement that the quality measurements are reliable, i.e., unbiased and sufficiently accurate. If the units of production are highly variable, e.g., many projects that save a little energy and a few that save a lot of energy, then simple random sampling, may not be cost-effective, i.e., it requires measurement of too many units.

Can you have a real-time process that accomplishes both evaluation goals or do you need separate evaluation processes? If both cannot be achieved by the same process, is the second goal sufficiently important in the long run? A manufacturer might rightly feel that careful real-time measurement will ensure that future production has sufficient quality. This may be a safe conclusion when measurements are automated, repeatable, and demonstrably reliable. In the more complex case of technical services, provided by humans, such as the development and delivery of energy efficiency projects, it may be harder to prove that only satisfying the first goal is sufficient.

3.1. Opportunities for Real-Time Evaluation

Real-time evaluation could theoretically occur at any of the following points in the delivery of efficiency projects. Some of these points occur prior to the program savings claim. The balance occurs after the savings claim. We describe information that could be gathered at each of these points to improve the program or that would contribute to the evaluation of savings.

1. **Pre-claim:**
    a. **Measure identification**. Is the measure identified appropriate for the application and correctly specified?
    b. **Baseline determination**. Is each measure associated with the correct RTF baseline (current practice or pre-conditions)?
    c. **Baseline data collection**. Has sufficient data been collected to reliably characterize the baseline, including sub metering of affected systems and equipment? For current practice baseline, this would include information on relevant codes, standards and end user practices. In the case of pre-conditions, this would include information about the likely remaining useful life of the affected systems, equipment and practices.

[3] Take the recent case of Volkswagen, where the number of poor quality units delivered to the market over a series of years has real consequences for the firm's current profitability. Further, independent testing of products sold, such as the tests performed by Consumer's Reports, e.g. levels of owner satisfaction with a prior model year, may influence the current and future demand for the manufacturer's products.

d. **Expected measure savings**. Has sufficient data been gathered about the efficient systems, equipment or practices to support a reliable estimate of likely savings?

e. **M&V plan**. All BPA efficiency projects require an M&V plan that describes what data to gather and how to estimate savings. Projects with expected savings greater than 200,000 kWh generally are required to collect more detailed site-specific data. The evaluation could determine whether the model used for estimating savings has the correct specification and whether there will be sufficient data after project completion to reliably estimate savings.

f. **Measure delivery and commissioning**. Were the projects implemented as planned? Are they operating as expected?

2. **Post-claim:**

a. **First year project savings**. Are the savings estimates reliable? Did they utilize all the available data regarding baseline and as-delivered performance? Did they consider any changes that occurred in the first year following project delivery?

b. **Lifetime project savings**. Do savings persist after the first year? Are the estimates of effective useful life, remaining useful life and typical annual savings reliable?

3.2. Summary of National Experience

We found only one example of a custom program evaluation that relies *exclusively* on real-time methods (evaluation of National Grid's New York Commercial and Industrial program, see Appendix B), and this is a post-claim design. In addition, we found that evaluators are conducting pre-claim reviews of selected custom projects for both NYSERDA and the CPUC. However, in both jurisdictions, the pre-claim review projects are still subject to sampling as part of the post-claim evaluation design. The pre-claim reviews are conducted in order to increase the likelihood that the ultimate evaluation estimates will closely correspond to the program's claimed savings. These reviews focus on projects with large expected savings, and for the CPUC, some with small saving that involve frequently delivered measures.

Based on comments from our expert panel and commentary from recent CPUC proceedings, we understand that the CPUC process has been contentious and time-consuming. Considerable time is spent on issues such as the likelihood of a project being a free-rider and the appropriate baseline. Even with extensive pre-claim review, the CPUC still reserves the right to revisit all modelling and assumptions about each project when evaluating the post-claim sample. This increases the contentiousness of the overall process. Some of the time spent may be devoted to debates over questions of policy, which could be resolved at a programmatic level instead of project-by-project. The NYSERDA process does not appear to be as contentious.

Regardless of how well a pre-claim review proceeds, the important finding is that neither of these two jurisdictions believe it is a substitute for evaluating the entire program after saving have been claimed. These early reviews could reduce the effort needed when the projects are sampled again, post-claim. However, both jurisdictions believe that the as-delivered features of

a project will still have to be confirmed and could be different from what was documented in the pre-claim review.

3.3. Benefits of Real-Time Evaluation

Real-time evaluation may provide the following benefits.

0.0.

1. **Continuous Program Improvement**. Projects can be sampled at any point during program delivery. Evaluation results can be shared with program staff on a project-by-project basis. For example, if the program introduced a new way of identifying efficiency measures for a certain class of customer, the first instances of this new approach could be sampled and examined to determine whether that innovation was effective.
2. **Improved Access to Baseline Data**. If projects are sampled prior to the project delivery there may be opportunities for additional baseline data collection, which would increase the reliability of savings verification once the project is complete.
3. **Evaluation Becomes Routine**. Program staff will consider evaluation to be a routine part of program operations if sampling occurs frequently throughout the program delivery process. This may reduce the effort required to educate and involve staff compared to post-claim designs that ask for input on a large sample all at one time.

3.4. Risks of Real-Time Evaluation

However, the following risks must also be considered.

0.0.

1. **Biased Realization Rates**. There is a risk of bias when real-time sampling occurs pre-claim and the evaluation provides feedback to the program as each project is reviewed. It is possible, that the feedback on a sample would be used by the program to improve all projects. However, that could only be proven by evaluating another sample selected post-claim, as is done by the CPUC and NYSERDA. Although not twice the effort, as much is learned from the pre-claim review of a project, sampling projects both pre- and post-claim will be more expensive and intrusive than the conventional design of selecting a sample once after all claims have been made for a period of program operation.

   This bias can be mitigated by sampling real-time, but post-claim. This is the design used in National Grid's current New York State Commercial/Industrial evaluation. This design provides feedback to the program each quarter, along with a true-up for the entire year of program operation. As part of this evaluation, the inspection of some projects might occur less than one year after a project is commissioned. This may introduce some challenges in evaluating first year savings as the conditions throughout that entire year may not be known at the time of the inspection.

   Another possible mitigation strategy, not yet tested, is to withhold feedback to the program until after savings claims have been made for a period of program operation. Projects could be sampled pre-claim, which would allow the evaluation to collect additional baseline data. The program would have to develop its savings claim without any reference to this

additional evaluation data. If this was achieved, the evaluation could compare its post-claim savings estimate to the program claim just as it would in a post-claim evaluation design. In addition to improved baseline data, the benefit of this design is that projects are sampled only once. This design should also speed up the evaluation process compared to the conventional post-claim design where a single sample is selected after all claims are made for a period of operation.

2. **Increased Sample Size**. The site-specific savings portfolio contains projects whose savings range from a few thousand kWh to a few million kWh. Simple random samples of such variable populations are inefficient. It might take a *simple random* sample with hundreds of projects to precisely estimate the mean savings per project. In a post-claim evaluation, project tracking data for all projects completed in a period is available and a *stratified random* sample can be selected. In a stratified design, projects with large savings will have a much greater chance of being selected. A stratified design can reduce sample size by a factor of 10 compared to a simple random design. Real-time evaluations require that projects be sampled throughout a period of program operation. A simple random selection can be used, but that will require a large sample. The alternative is to guess at the number of projects that will be completed in the entire period and where any one project will fall in the distribution of project savings. Some efficiency can be gained by good guess work, but it will come at the cost of greater complexity in managing the sampling process.

   In some cases, a project sampled real-time before it is completed may be a dry-hole, i.e., never completed. Accommodating the dry-hole rate can further expand the real-time sample size. Some real-time designs may also sample for specific periods, for example by quarter. This imposes further stratification by time which may also increase sample size.

   Real-time pre-claim sampling, when feedback is provided, may reduce the variance in evaluated saving realization rates. This reduced variance could reduce the size of the sample required to achieve a target sampling precision. However, this only affects the size of the post-claim sample, and requires sampling both pre- and post-claim as is done by the CPUC and NYSERDA to avoid introducing bias.

4. BPA's Real-Time Evaluation Options

In this section, we provide strategic advice regarding BPA's options for implementing real-time evaluation. In order to implement a real-time evaluation, BPA must select the appropriate evaluation design, determine the role that third-parties will play in implementing that design and consider how that design will be implemented for each of its program delivery channels.

4.1. Possible Evaluation Designs

This research has identified four possible evaluation designs. Two of these are real-time designs that have been or are being implemented in other jurisdictions. One is a design that was suggested by BPA and was discussed with members of our expert panel. The final one is the conventional design implemented by BPA for its 2012-13 site-specific savings portfolio,

which was not a real-time design, but is described here so that it can be easily compared to the real-time options.

0.0.

1. **Real-time Post-Claim**. This design is currently being implemented in the evaluation of lighting projects by National Grid in New York State. Stratified post-claim samples of projects are selected following the close of each calendar quarter. Feedback can be provided to the program based on each of these quarterly samples, without increasing the risk of bias as projects are selected post-claim. Even if the program makes changes based on the feedback, the sample in subsequent quarters will account for these changes. However, caution should be exercised in releasing quarterly feedback as the program and others may expect the results from one quarter to be representative of the entire program period. Regardless of whether quarterly results are shared, this design should reduce the average time between project completion and completion of the evaluation for a period of program operation. Projects can only be sampled once under this design. This design may increase sample size, as the stratification involves some guesswork about the size distribution of the projects to be completed in any period. This design does not provide any opportunity for additional baseline data collection by the evaluation, thus it may increase the risk of measurement bias, relative to the designs that involve pre-claim sampling.
2. **Real-time Pre- and Post-Claim**. This is the design currently used by the CPUC and NYSERDA. This design involves real-time pre-claim sampling to provide rapid feedback to the program. It also may result in improved and less variable savings realization rates, thus reducing the sample size for the post-claim sample. The post-claim sample can be efficiently stratified as it is not drawn until all savings claims are made for a period of program operation. If all projects are subject to the post-claim sampling, this design does not increase the risk of sampling bias. The real-time pre-claim sample can be focused on projects with large savings or projects with small savings that involve frequently delivered measures. Some projects may be sampled twice. This is especially likely for projects with large savings. This design provides opportunities for additional baseline data collection by the evaluation for projects in the pre-claim sample, thus decreasing the risk of measurement bias. Yet, this approach likely increases total evaluation effort.
3. **Real-time Pre-Claim, No Feedback**. This design was suggested by BPA and was discussed with members of our expert panel. We do not know of any jurisdiction where it has been tried. This design involves real-time pre-claim sampling on a random sample of projects. This design allows the evaluation to collect additional baseline data. This additional data may improve the reliability (decrease measurement bias) of estimated savings, which are developed by the evaluator after the program claims savings for the projects. In addition to improved baseline data, the benefit of this design is that projects are sampled only once. This design should also speed up the evaluation process compared to the traditional post-claim design where a single sample is selected after all claims are made for a period of program operation. This design may increase sample size as stratification will involve some guesswork about the size distribution of the projects to be completed in any period and there may be lost projects due to non-completion. To minimize potential for bias, the program would have to develop its savings claim without any reference to this additional

evaluation data and the evaluation would need to withhold any feedback on its findings on specific projects until savings claims have been made for a period of program operation.

4. **Traditional Post-Claim Sampled Once**. In this design, sampling does not occur until after the end of a period of program operation, when savings for all projects have been claimed in the BPA reporting system. This is the design used in the 2012-13 portfolio evaluation. Efficient sample stratification can be applied as all claims are available when the sample is drawn. Projects will only be sampled once. This design has the longest delay between project completion the completion of the evaluation. It does not increase the risk of sampling bias. However, like the other design that only involves post-claim sampling, there may be measurement bias due to the limitations on the collection of baseline data. Efforts can be made to re-construct baseline conditions from documentation gathered by the program staff, inspection of the affected systems and interviews with operators and tenants, but there are always uncertainties in such re-constructions. This is the most common evaluation design used by other utilities and regions.

4.2. Role of Third Parties

Third-party, contracted, independent evaluators could play a role in implementing any of the evaluation designs described above. Following are general options for possible roles for third-parties; detailed design work would be required for any specific program evaluation.

0.0.

1. **None**. All evaluation work could be accomplished by program staff as part of their quality control activities. Staff could be organized so that no one evaluates their own work. We have not seen any instances of other utilities or regions conducting evaluation in this manner, and this approach violates the national consensus definition of evaluation. Alternatively, BPA could have the evaluation performed by members of its staff that are in a separate evaluation department. As long as this staff does not have responsibilities for program operation and is not managed by those who operate the program this would be consistent with the national consensus definition of evaluation. There are many examples of such staff managing evaluations, performing quality control on evaluations, and conducting all or a portion of entire evaluations. Third-parties most frequently perform most of the evaluation work because utilities and other agencies do not have sufficient staff to complete all the required evaluation work.

2. **Some: Protocols, Sampling, Review**. Third-parties could be retained to specify the evaluation protocols, including sampling, model selection, and data collection. They could also be retained to review a sample of projects for the purpose of determining whether the protocols had been followed. This could be done real-time without creating any additional data collection burden on the program staff or end users. The national consensus definition of evaluation is not clear on which party must collect the data used by the evaluation. In the traditional designs, because they sample only after the savings claim, the evaluators mostly rely on baseline data collected by the program. However, in those designs the evaluator has the opportunity to collect their own post data and to confirm some of the baseline data. In our opinion, if the third-party agrees to the data collection and savings estimation

protocols, and has the opportunity to confirm that these protocols are followed; this role would be consistent with good evaluation practice.

3. **Full evaluation**. Third-parties could play the same role that they did in the 2012-13 portfolio evaluation and be responsible for all aspects of the evaluation. This may impose additional data collection burden on the program staff and end users. This is consistent with the definition of evaluation.

4.3. Program Delivery Channels

BPA's site-specific portfolio is delivered through three channels. Each of these present different challenges for evaluation. It may not be possible to implement all of the evaluation designs described above for each of these channels. These channels and the challenges for each are as follows:

0.0.

1. **BPA Staff Operated**. This channel delivers custom projects to commercial, agricultural and federal end users served by Option 1 utilities. This work is primarily accomplished by members of BPA's engineering staff although some assistance is provided by the staff of Option 1 utilities and BPA staff may choose to involve outside vendors in various phases of the work. Quality control and M&V is carried out by BPA staff.
2. **Program Implementation Contract**. This channel delivers projects to industrial end users, via a third-party program operator. The program operator provides marketing, project development, M&V and quality control services for end users of Option 1 utilities and some of these services to selected projects of Option 2 utilities, with the authorization of those utilities.
3. **Option 2 Utility Operated**. Option 2 utilities are responsible for this channel, although they operate under the requirements of BPA Implementation Manual, which by reference includes an expectation of compliance with BPA M&V protocols. As BPA lacks visibility into pre-claim information, additional cooperative agreements would be needed to implement real-time pre-claim sampling.

5. M&V 2.0 IS A SEPARATE CONCEPT

M&V 2.0 has received a lot of national attention in the last few years. It generally refers to the use of regression models of interval metering data (hourly or less[4]) to estimate whole-building energy savings. If sub metering is available these techniques can also be used to estimate savings for specific building systems. Some utilities have implemented Advanced Metering Infrastructure (AMI) that allows for collection of interval data from all of their end users. This opens the possibility of implementing pre/post evaluation designs that only require data which the utility is already collecting for billing purposes along with publicly available weather data.

[4] Although AMI data may be collected for 15 or 30 minute intervals, it is often aggregated to the daily level in order to reduce serial correlation in the regression models.

For some utilities, it is also possible that this modelling could be automated so that it could be carried out for all efficiency projects, avoiding the problems with real-time sampling described above. Unfortunately, there are a number of challenges to the M&V 2.0 vision, including non-routine adjustments to building operations, the treatment of current practice baselines, and determinants of use other than weather, e.g., production level. All these make it difficult to estimate savings based solely on weather and interval billing data. BPA faces the further challenge that only a portion of its customer utilities have AMI and BPA lacks direct access to customer billing data.

M&V 2.0 requires both pre- and post-billing data to estimate savings. Savings estimation can start shortly after a project is complete. However, a reliable estimate of first-year savings generally requires months of post billing data[5]. In some cases, a full year of data is required. If this technique is applied to all projects, feedback to the program could begin earlier than would occur using a post-claim evaluation design. However, even in that case, the estimation of savings for a program period requires estimates of savings from all projects completed in that period. If a year of post data is needed for the last projects completed, this process could take longer than a post-claim design that relies on engineering models to estimate savings. In addition, M&V 2.0 techniques provide only limited feedback during the project development (pre-claim) portion of the program process. Unless they are built on sub metering data, they may also provide limited feedback after completion on savings for specific measures.

[5] This is true regardless of when the project is sampled. Real-time sampling can accelerate findings for individual projects, but you still have to wait for post billing data to accumulate before modeling any particular building. If there is interval metering, preliminary models can be built and may be useful in tracking or diagnosing building performance, but substantial post period data is needed before an accurate estimate of annual savings can be derived.

A. ANNOTATED CITATIONS - DEFINITION OF EVALUATION

SBW conducted a search of the following websites looking for papers relevant to the definition of evaluation for custom programs.

- ACEEE.org (American Council for an Energy-Efficient Economy)
- CPUC.ca.gov (California Public Utilities Commission)
- ASE.gov (Alliance to Save Energy)
- EPA.gov (U.S. Environmental Protection Agency)
- EMP.lbl.gov (Electric Market and Policy)
- Energy.gov (U.S. Department of Energy)
- RTF.nwcouncil.org (Regional Technical Forum)
- EVAL.org (American Evaluation Association)

On each of the websites, we searched throughout the relevant tabs, listings, databases, links, publications, records, presentation, and conference summaries. Generally, we used the keywords "energy efficiency evaluation," "energy efficiency measurement and verification," "industrial energy conservation projects" and "impact evaluation." We also asked Dr. Richard Ridge to cite other sources that were relevant to this memo.

Relevant portions of the papers and reports discovered in this search are provided below.

0.0. **1.** *The California Evaluation Framework,* TecMarket Works Framework Team, June 2004 (Latest Revision January 2006)

The California Evaluation Framework (Framework) provides a consistent, systemized, cyclic approach for planning and conducting evaluations of California's energy efficiency and resource acquisition programs. This document presents that Framework and provides valuable information concerning when evaluations should be conducted, the types of evaluation that can be conducted, and a discussion of approaches for conducting those studies. The intended audience for various sections includes policy staff, program portfolio managers, program planners and implementers, evaluators, and other stakeholders. Credibility of evaluations and evaluators are absolutely essential for evaluations to fill their role in providing findings on the results from the program and for providing recommendations for program refinement and investment decisions. The Framework includes the American Evaluation Association (AEA)'s set of guiding principles for evaluators and recommends that these principles guide energy program evaluations and the evaluators who conduct these studies.

2. *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*, TecMarket Works Team, April 2006

This document is to be used to guide the efforts associated with conducting evaluations of California's energy efficiency programs and program portfolios. The Protocols are the primary guidance tools policy makers will use to plan and structure evaluation efforts and that staff of the California Public Utilities Commission's Energy Division and the California Energy

Commission, and the portfolio (or program) administrators will use to plan and oversee the completion of evaluation efforts. The Protocols are also the primary guidance documents evaluation contractors will use to design and conduct evaluations of programs. The Protocols define the skill sets required for various impact evaluation tasks in the Impact Protocol but do not otherwise define evaluators.

3. *Impact Evaluation Framework for Technology Deployment Programs*, John H. Reed (Innovologie LLC), Gretchen Jordan (Sandia National Laboratories), Edward Vine (Lawrence Berkeley National Laboratory), July 2007

The impact evaluation framework is specifically designed to assist energy program managers and evaluators in Federal, state, and local governments and in public entities and institutions that are increasingly accountable for delivering and demonstrating results. The framework provides a series of steps and some templates that evaluation contractors and program managers can use to develop meaningful impact evaluations that help refine their programs, increase program effectiveness, make the tough decisions to drop ineffective program elements, and develop credible evidence to help communicate the value of the program to stakeholders. This document is less about measurement and analysis techniques and more about providing tools that focus on defining the linkages between outputs and outcomes. The idea is to use sound principles of social science to more clearly identify what needs to be measured, develop better evaluation designs, and better harness existing data collection activities to obtain needed data. There is one reference to evaluators and program managers with regard to avoiding potential conflicts of interest by maintaining some degree of separation throughout the process. This creates an extra burden on the part of programs to create transparent and defensible evaluation processes and conduct quality evaluations using independent evaluators.

4. *Model Energy Efficiency Program Impact Evaluation Guide*, Steven R. Schiller (Schiller Consulting, Inc.), December 2007

This Guide describes a structure and several model approaches for calculating energy, demand, and emissions savings from energy efficiency programs. By adhering to best practices and standard procedures, stakeholders can use program evaluation as an effective tool to support the adoption, continuation, and expansion of energy efficiency programs. Chapter 7 builds on preceding chapters and presents the steps involved in planning an impact evaluation. Either the program implementer or a third party typically conducts the evaluation. The third party—valued for a more independent perspective—can be hired either by the implementer, with criteria for independence, or by an overseeing entity such as a utility regulator. A typical approach for utility-sponsored efficiency programs is for the utility's evaluation staff to manage studies that are completed by third-party consultants, whose work is reviewed by the utility regulatory agency. The objective is for all parties to the evaluation to believe that the reported results are based on valid information and are sufficiently reliable to serve as the basis for informed decisions. There are advantages and disadvantages to using either implementers or independent third parties to conduct evaluations—selecting one or the other depends on the goals of the evaluation. Regulated energy programs and programs with a financial outcome hinging on the results of the evaluation tend to require third-party evaluation. Another

approach is to have the evaluation completed by the implementer with the requirement for third-party verification.

5. *Energy Efficiency Program Impact Evaluation Guide*, SEE Action, EM&V Working Group, April 2011

The Energy Efficiency Program Impact Evaluation Guide describes and provides guidance on approaches for determining and documenting energy and non-energy benefits resulting from end-use energy efficiency programs and portfolios of programs. It specifically focuses on impact evaluations for programs designed to reduce facility energy consumption and/or demand as well as related air emissions. The Guide provides definitions for the roles of principals in an evaluation process, administrators, implementers and independent third-party evaluators. The designation of "independent" and "third-party" is determined by those entities involved in the use of the evaluations and may include evaluators retained, for example, by the administrator or a regulator. Clear definition for the relative roles of the administrator, implementer, and independent third-party evaluator is an important activity of the planning process.

6. *National Energy Efficiency Evaluation, Measurement and Verification (EM&V) Standard: Scoping Study of Issues and Implementation Requirements*, Steven R. Schiller (Schiller Consulting, Inc.), Charles A. Goldman (LBNL Environmental Energy Technologies Division), Elsia Galawish (iTron), EM&V Working Group, April 2011

This paper, funded by the DOE, is a scoping study that identifies issues associated with developing a national evaluation, measurement and verification (EM&V) standard for end-use, non-transportation, energy efficiency activities. The objectives of this study are to identify the scope of such a standard and define EM&V requirements and issues that will need to be addressed in the course of developing such a standard. With regard to defining the role of the evaluator in such a broad context, the paper proposes the evaluators should ideally be impartial in their work and not have their compensation tied to the magnitude of their impact evaluation results. However, in many states, energy efficiency program administrators often fulfill many EM&V roles (for cost savings or other reasons). Thus, as part of developing a national EM&V protocol, it is likely that concepts such as "independent evaluation" and/or "third-party evaluation" will need to be defined. It is possible that acceptable institutional models and arrangements for what organizations or types of organizations conduct various types of EM&V activities may also have to be discussed.

7. *Meaningful Impact: Challenges and Opportunities in Industrial Energy Efficiency Program Evaluation*, Anna Chittum, September 2012

Based on interviews and surveys with industrial energy efficiency program managers, evaluators, and regulators, this report discusses how industrial energy efficiency program evaluation is conducted and the types of data and metrics derived by evaluators. It explains the use of these various metrics and the manner in which specific metrics are developed. The paper defines an evaluator as, "An individual or organization tasked with the evaluation of an energy efficiency program. Most often, this is not a member of the organization administering the energy efficiency program, although internal evaluators within program-administering entities such as utilities and public benefit fund organizations do exist." The paper indicates that external evaluators perform 85% of industrial evaluations, while internal evaluators of the

funding organization conduct 11%, and 4% are internal evaluators at regulatory agencies. Major evaluation activity is largely the domain of third-party consultants, and the use of third-party consultants to conduct evaluations is common among all types of efficiency programs. Third-party consultants share best practices and are actively involved in a number of different groups to help maintain the freshness of their approaches and to learn about new techniques and technologies. However, some staff at industrial energy efficiency programs noted that they felt it was critically important to have their staff more involved in the evaluation process to make sure evaluators fully understood the program's internal data and practices. Among industrial programs, there is considerable variation in the degree to which program staff is involved in evaluation activities.

8. *Project Manager's Guide to Managing Impact and Process Evaluation Studies*, Yaw O. Agyeman (Lawrence Berkeley Laboratory), Harley Barnes (Lockheed Martin), August 2015

The purpose of this Guide is to help managers of the DOE's Office of Energy Efficiency and Renewable Energy manage evaluation projects with the goal to create and manage objective, high quality, independent, and useful impact and process evaluations. The step-by-step approach described in this Guide is targeted primarily towards program staff with responsibility for planning and managing evaluation projects for their office, but who may not have prior training or experience in program evaluation. The objective is to facilitate the planning, management, and use of evaluations. While the definition of an evaluator evidently is one who executes the evaluation process as outlined in the Guide, it does unequivocally refer to the evaluator as "independent external expert evaluators," who are generally selected through a competitive bidding process.

9. *Guidelines for the Estimation of Energy Savings*, Northwest Power and Conservation Council, Regional Technical Forum, December 2015

The purpose of this document is to describe how the Regional Technical Forum (RTF) selects, develops and maintains methods for estimating savings from the delivery of energy efficiency measures. Four savings estimation methods are defined: Unit Energy Savings (UES), Standard Protocol, Custom Protocol and Program Impact Evaluation. It is the RTF's intention that each method will produce savings estimates of comparable reliability sufficient to meet the needs of regional energy planners. These methods are also expected to support regulatory processes related to the adoption and planning of energy efficiency programs. The Guidelines state that impact evaluations should be designed to achieve reliable estimates of savings while accommodating the special requirements of the program's delivery methods, target markets, efficiency measures, operating agency, and regulatory environment. No guidance is provided on whether the impact evaluator should be a third party or otherwise. The Guidelines do provide recommendations on the evaluator's skills, such as selecting and designing samples to maximize the precision of a given sample size, estimating savings using a variety of engineering and statistical techniques, and to extrapolate sample findings to the study populations and to quantify the uncertainty in this extrapolation.

10. *Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*, US Department of Energy, National Renewable Energy Laboratory.

Under the Uniform Methods Project, the DOE is developing a set of protocols for determining savings from energy efficiency measures and programs. The protocols provide methods for evaluating gross energy savings for residential, commercial, and industrial measures commonly offered in ratepayer-funded programs in the United Sates. The measure protocols are based on a particular IPMVP option, but provide a more detailed approach to implementing that option. Each chapter has been written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The definition of an evaluator in the 2012 SEE Action Guide was considered sufficiently detailed to be adopted by the Uniform Methods Project (see Chapter 1: Introduction). However, the UMP doesn't provide the additional detail provided by the SEE Action Guide that "… in general practice 'independent third party' is thought to mean that the evaluator has no financial stake in the evaluation results."

B. Annotated Citations - Real-Time Evaluation

We conducted a search of the following websites looking for evidence of real-time evaluations for custom efficiency programs. We focused our search on the time period of January 2010 thru July 2016.

- ACEEE.org (American Council for an Energy-Efficient Economy)
- IEPEC.org (International Energy Program Evaluation Conference)
- CPUC.ca.gov (California Public Utilities Commission)
- PSE.com (Puget Sound Energy)
- UTC.wa.gov (Washington Utilities and Transportation Commission)
- CALMAC.org (California Measurement Advisory Council)
- NEEP.org (Northeast Energy Efficiency Partnerships)
- library.CEE1.org (Consortium for Energy Efficiency) Energy Efficiency Program Library
- RTF.nwcouncil.org (Regional Technical Forum)
- ma-eeac.org/studies (Massachusetts Energy Efficiency Advisory Council)
- www.energizect.com/connecticut-energy-efficiency-board/about-energy-efficiency-board/library (Connecticut Energy Efficiency Board)

On each of the websites, we searched throughout the relevant tabs, listings, databases, links, publications, records, presentation, conference summaries, etc. Generally, we used the keywords "real-time evaluation," "custom evaluation," and "concurrent evaluation;" and we also made use of just plain old browsing, clicking, clicking some more, and poking around. Similarly, we searched through the websites of several energy efficiency firms. We also asked and expert panel consisting of Pete Jacobs, Jon Maxwell and David Jacobsen to cite other relevant sources.

Relevant portions of the papers and reports discovered in this search are provided below.

1. It's About Time: Doing Integrated Real-Time Impact Evaluations, Sue Haselhorst, ERS, and Joe Dolengo, National Grid, ACEEE Summer Study 2016

This paper describes the evaluation of National Grid's New York Commercial and Industrial program. The evaluation was partially complete as of the publication of this paper, having reported on results from samples drawn in two consecutive calendar quarters. The evaluation is designed to provide quarterly results from measurement and verification, along with process-oriented feedback. Completed projects are sampled shortly after the end of each calendar quarter.

2. *THE CHANGING EM&V PARADIGM: A Review of Key Trends and New Industry Developments, and Their Implications on Current and Future EM&V Practices*, DNV GL, December 2015

This paper focuses on new and evolving analytic tools and methods that provide automated, ongoing analysis of energy consumption data and how this impacts M&V and EM&V practices. This paper characterizes the trends in data collection and analysis with the purpose of furthering stakeholders' understanding of how these approaches can be leveraged for EM&V, as well as discussing the limitations of these new tools and techniques. The research examines how and to what extent the enhanced data and new tools can help address stakeholders' concerns that standard EM&V procedures are costly and results take a long time to produce. Similarly, the paper examines whether these new capabilities can maintain or improve the accuracy and reliability of EM&V.

3. *Evaluation and Regulatory Teamwork: Closing the Custom M&V Gap*, Kris Bradley (Itron), Kay Hardy and Peter Lai (CPUC), August 2015.

This paper examines an on-going multi-year improvement process for custom energy efficiency projects and programs targeted at the non-residential sector. The process uses a combination of policy guidelines, ex-ante review, program requirements, ex-post evaluation, and QA/QC procedures to improve both custom impact estimates and custom incentive programs. This paper addresses incentive programs in California that focus on custom offerings. The anchor to this effort consists of a relatively new ex-ante review (EAR) activity. EAR involves the parallel participation by CPUC staff and their (evaluation) contractors in the review of ex-ante savings estimates, providing guidance and recommendations to the programs for a sample of selected projects. EAR is an evaluation-oriented regulatory approach that could potentially be applied in other jurisdictions where custom energy efficiency programs operate, in an effort to improve program processes and procedures and evaluation results.

4. *Leaving the Rearview Mirror Behind: Assessing the Effectiveness of a Concurrent Impact Evaluation Process*, Betsy Ricker and Nick Collins (ERS), Cheryl Glanton and Carley Murray (NYSERDA), presented at IEPEC August 2015.

NYSERDA's Industrial and Process Efficiency (IPE) program implemented a concurrent evaluation process in which impact evaluators worked alongside program implementers on complex projects with high expected savings that were considered to have a potentially high risk for significant differences between realized and predicted savings. Evaluators provide early feedback to the program staff on key evaluation variables, baseline characterization, and M&V methodologies. The process resulted in increased realization rates and reduced the number of

required end user contacts. Costs were higher than for traditional evaluation processes. Although evaluators conducted an independent analysis, the process could potentially compromise the evaluator's independence. Others implementing this process should be prepared to be flexible as the process will evolve over time.

5. *Pre-Retrofit Evaluation of Industrial Projects***,** Jonathan B. Maxwell and Betsy Ricker (ERS), Carley Murray (NYSERDA), August 2013.

Through this pre-retrofit review process, NYSERDA's impact evaluation contractor works with program implementation staff prior to measure installation to review a sample of the large and complex projects, particularly those that involve process-specific baseline definition. The evaluators' pre-installation activities include site visits, review of savings calculations, writing early evaluation assessment reports, and periodic meetings with program implementation staff and NYSERDA's technical review contractors. This paper shares lessons learned and analytical techniques used to ensure that evaluators and the program team gain the benefits of increased engineering rigor and higher savings realization rates.

6. *Learning from Public Health: Embedded Evaluation and its Applications to Energy Efficiency*, Courtney E. Henderson and Anne Dougherty (ILLUME Advising), August 2015.

This paper reports on how the public health sector adopted real-time evaluations concurrent with program planning, rollout, and operation in the mid-1990's and asserts that energy efficiency field could benefit from a similar approach. The paper emphasizes the importance of broadening the stakeholder group to include end users, funders, and regulators alongside program teams and administrators. Stakeholders are kept informed by the evaluators throughout each stage of the evaluation and, in particular, their participation is critical at the evaluation planning and design stage. Evaluators provide program teams and end users with tools to assess their own outcomes on an ongoing basis in support of the evaluation. Further, evaluators provide timely feedback to all key stakeholders to share results and inform recommendations for future program design and provide data on what's working well and why. If the primary goal of evaluations is to inform and improve programs, this approach can correct ineffective strategies before significant resources are spent.

7. *Industrial and Process Efficiency Program Impact Evaluation (2010-2012)*, prepared for NYSERDA by ERS, April 20, 2015. Sections 3.2, 4.3, and Appendix D.

This is the final report associated with the Ricker paper above. It contains comparative results based on concurrently reviewed 10 projects. The content includes assessment of RRs for concurrently reviewed projects compared to non-concurrent projects. Most were ex post evaluated at 3% to 6% of the ex ante estimate, a tighter range and closer to 1.0 than projects without concurrent evaluation.

8. *Ex Ante Review Fact Sheet #2: The Commission's Ex Ante Review Process*, California Public Utilities Commission, June 19, 2014.

This fact sheet describes the history of, purpose and process of conducting ex ante review in California for custom projects and the Database of Energy Efficiency Resources (DEER).

9. *How Information and Communications Technologies Will Change the Evaluation, Measurement, and Verification of Energy Efficiency Programs*, Ethan A. Rogers, Edward Carley, Sagar Deo, and Frederick Grossberg, American Council for an Energy Efficient Economy, December 2015.

The past decade has seen the development of a large number of technologies that provide significant opportunities for real-time data collection by EM&V practitioners. This paper provides a broad overview, not specific to any particular evaluation, of technologies such as smart utility meters, cloud-ready smart thermostats, network ready HVAC equipment, remote-access building analysis software, management and process control systems, cloud computing, and remote analytics - all offering new or improved capabilities for gathering and analyzing energy data. This field is evolving very quickly with many enhanced capabilities available now and others coming online in the near term.