Pilot Programs to Mainstreaming: How Evaluation Can Help

Linda Dethman, The Cadmus Group, Portland OR
Shahana Samiullah, Southern California Edison, Rosemead CA
Anne West, The Cadmus Group, Portland OR

Abstract

In the United States many organizations, including those that sponsor energy efficiency programs, conduct pilot or trial programs to help them decide whether program ideas are good enough to be rolled out on a larger basis either as part of existing programs or as new programs. Despite the allure of pilot programs, however, we found consistent evidence that they often do not provide the needed guidance for the road ahead. We suspect that the root of this problem lies with the intersection of pilot design and evaluation approaches.

Thus, the goal for this paper is to provide program designers, implementers, and evaluators with insights about setting up and conducting better evaluations of energy-efficiency program pilots. This paper first presents two useful, but not often used, frameworks for conducting pilots–experimental and quasi-experimental designs. It then explores two pathways to assess energy-efficiency pilots: (1) involvement at pre-launch, where evaluation is incorporated into the fabric of the pilot; and (2) involvement at post-launch, where evaluation activities are designed after the pilot is underway.

To illustrate key evaluation challenges and solutions of evaluating pilots, we then present experiences with three pilot programs in the United States: an LED pricing trial, an appliance recycling retailer trial, and a pilot that encourages behavioral changes to reduce energy use in the workplace. These examples describe issues associated with the evaluation design and development that evaluators typically encounter when asked to assess pilots. Finally, we provide a systematic approach for evaluators to use when asked to assess pilot efforts, no matter what state they find them in. (Note that in this paper, we use the term *pilot* and *trial* program interchangeably.)

What Are Pilot Programs and Why Are They Challenging to Evaluate?

Energy-efficiency programs in the United States are designed to meet certain overall goals, such as acquiring resources and transforming markets. Other federal, state, or local goals (such as economic development or equitable services for all consumers) may also be part of these programs. Since investment in energy-efficiency programs can be substantial, and the viability of many elements of programs may be uncertain, program sponsors—such as utilities—often devise pilot programs. These pilots test new program ideas or new elements of existing programs in new settings before deciding whether to launch the approaches on a wider basis.

While the idea of pre-testing a program is appealing, pilot programs differ somewhat from full-scale programs in that they: (1) usually operate within a more constrained set of circumstances; (2) are delivered and evaluated within a relatively short time frame; and (3) may contain features or be subject to outside influences—such as higher incentives, greater support from sponsors, or changes in policy—that may not be carried forward into full-scale programs.

As with a full-scale program evaluation, a pilot program evaluation examines assumptions and methods, measures how well the pilot meets its goals (both in terms of its process and energy impacts), and recommends program improvements. Unlike evaluations of full-scale programs, however, an evaluation of a pilot program needs to provide a roadmap for the future of a program

concept that often has been implemented on a limited, short-term basis and with specialized features.[1]

Because pilots are trials (or "proof of concept" programs), intended for possible mainstreaming,[2] they necessitate more, rather than less, information about the problem, the solution, the market, implementation, and performance. Thus, the nature of many pilot programs is likely to make them both especially important—and more challenging—to evaluate. The evaluation efforts may also be relatively more costly (as a proportion of program costs), because data are limited, time is short, and results must be placed within what is possible for future action.

In the next sections of this paper, we describe and give examples of two common pathways to evaluating pilot programs: pilots where evaluation is part of the pilot design (pre-launch) and pilots where evaluation design mostly takes place after the programs are underway (post-launch). We also identify the challenges that both pathways may present to evaluators.

For both paths, we suggest evaluators conduct an evaluability assessment (EA, see appendix for additional discussion), and work with program designers, managers, and implementers. EAs are tools for designing a meaningful evaluation framework and data collection system so that the essential data are available for the evaluation. This tool assists evaluators in the following: (1) defining the problem and research objectives; (2) identifying the data that need to be collected; and, (3) specifying how it will be collected and used. Since an EA is useful as both a program planning tool and an evaluation tool, a separate section of this paper discusses the EA concept and applies it to pre-launch and post-launch pilot program evaluations.

Experimental and Quasi-Experimental Pilot Program Designs

In their most manageable and evaluable form, pilots are carefully planned projects that test a small number of well-defined concepts. Many pilots, however, are more complicated and have multiple moving parts and limitations, which, in turn, often present greater challenges for evaluators. All too often, pilots are not well designed, and are added as an afterthought alongside or within current programs. There may be too many areas and different elements trying to be tested in one pilot. In addition, with the limited sample sizes in pilot programs, it can be a challenge to define the different treatment and control groups and to draw valid conclusions when comparing them.

In this paper we focus on experimental and quasi-experimental designs for pilot programs, since we believe these frameworks, while by no means foolproof, help produce the best results for program planning. Both experimental and quasi-experimental designs provide a framework for evaluations, and each has its purposes and applications. A true experimental design can only be used in a pre-launch scenario; however, quasi-experimental designs (and their variations) can be used in pre-launch and post-launch pilot program evaluation approaches. A common approach we have seen for piloting energy-efficiency programs is to try out an idea or design but without having the control groups that experimental designs require. However, we have chosen to simplify our discussion by focusing on experimental and quasi-experimental designs for pilot programs.

Both experimental and quasi-experimental evaluation designs compare two groups to determine whether the desired program outcomes are more likely to occur in the intervention or treatment group. A familiar example of this is testing whether raising the incentive makes a difference in program participation. Thus, an experiment could be designed where the incentive level is increased in a treatment group, while the control group maintains a steady incentive level.

1 As Caruth and Bardeaux describe, ascertaining how to push a fledgling program—even when it isn't technically a pilot program—to its full potential can take concerted effort. In a second paper, co-authors Murray and Fawcett note that, "A change in government halfway through the pilot's delivery timeframe saw the policy landscape change. This presented clear risks to the perceived relevance and take-up of the evaluation findings." The authors of these two papers focus on how pilot program evaluations, used as a roadmap, can "…remain relevant and influential against the context of a changing policy landscape..." for a future program.

2 In the U.S., "mainstreaming" is a term of art, meaning the program is not a pilot, but is a fully developed program offered to eligible customers.

Developed before a program launches, an experimental design incorporates evaluation into the fabric of the pilot program. Experimental designs—long a staple of applied social science research—are used to establish a causal or correlational relationship between the program intervention and the measured outcomes. Essentially, experimental designs provide the framework within which the program operates.

In an experimental design pilot program, the treatment and control groups are randomly pre-assigned from either a randomly drawn population or from the program's target population. (Random assignment ensures that the groups are equivalent from a statistical point of view.) Through this random assignment, true experimental designs try to control as much as possible the number of non-treatment variables to obtain a better picture of cause and effect. However, because true experimental designs could create an artificial situation that will not occur under typical circumstances, the ability to generalize results to real situations is sometimes limited. In applied energy-efficiency program evaluation research, true experimental designs are often difficult to implement with multiple groups and test variables, can be challenging to control, and add cost to the experiment. [3]

Pilot programs can also employ a quasi-experimental design in either pre-launch or post-launch evaluations; this approach is more frequently seen in energy-efficiency pilot programs. In a quasi-experimental design, customers are not randomly assigned to a treatment or a control group. This is often because the program cannot restrict participation, since programs typically serve all eligible populations. The participants form the treatment group, while the comparison group members are typically matched to them on key characteristics. For example, the comparison group can be composed of program-eligible customers who don't participate, or, future participants.

The quasi-experimental pilot program faces challenges because the evaluator cannot control many of the factors that can affect the program results. However, a quasi-experimental design often makes generalization to the real world more apparent, so quasi-experimental designs tend to be used in more real-world applications where an experimental design cannot be used.

Designing Evaluation into Pilot Programs: Pre-Launch

Evaluations designed before a pilot program's launch tend to have more options around the evaluation approach and more flexibility in the design, since the evaluation can be incorporated into the structure of the pilot. Pre-launch evaluation planning offers an opportunity to select the most appropriate evaluation approach, which could be an experimental design, quasi-experimental design, or something else. The intent here is to design the pilot and the evaluation simultaneously in a way that facilitates evaluation.

Conducting an EA as the first step in developing an evaluation plan brings structure and assists staff in identifying and answering key questions. Clearly defining both the problem the pilot intends to solve and the program's objectives are the keys to structuring an evaluation that provides meaningful results regarding the pilot's success. Evaluations designed before a pilot launches provide opportunities to assess available data and collect any additional data needed to evaluate the pilot and provide clear results.

Designing Evaluation into Pilot Programs: Post-Launch

In many cases, evaluators are asked to assess pilot programs after they are already in motion or even after they have ended. Here again, an evaluability assessment can illuminate many aspects of the pilot, including its intent, purposes, desired outcomes, available methods, and data needs. This review will convey to the evaluator what is known, what is available, and where the gaps are. In some cases, it may be possible to impose a quasi-experimental design but in other cases the design

3 While an experimental design is the exception in demand side management (DSM) pilot evaluations, they are the current practice in critical peak pricing and real time pricing experiments.

will need to be built around available data. Still, after assessing what is possible, evaluators are typically able to design and implement some level of evaluation activities that will produce plausible, useful, and actionable results from existing sources, such as management databases, customer billing and consumption histories, and other systems data. In addition, it may be possible to collect primary data from pilot stakeholders and their counterparts in control groups.

Assuming that the pilot program has collected—or is designed to collect—the right amount and type of program information in its tracking systems, then an evaluation can still be done. A quasi-experimental design employing treatment and comparison groups may be the most appropriate evaluation approach; the post-launch section of this paper provides an example of this approach.

Pre-Launch and Post-Launch Pilot Program Examples and Their Challenges

This section provides three examples of pilots working, at least in part, within a quasi-experimental design framework. For two of these pilots, the evaluations were designed before the pilots launched; for the third pilot, the evaluation was designed well after the pilot launched. Issues affecting evaluation design and implementation are discussed for each example. These challenges and solutions are instructive and transferable to evaluations of other pilot programs.

Pre-Launch: Two Examples – An LED Pricing Pilot and an Appliance Recycling Retailer Pilot

The LED Pricing Pilot. Southern California Edison (SCE) implemented this pilot in its service territory as a quasi-experimental design. The primary purpose was to determine the best approaches for incentivizing LEDs in a residential program. The secondary purpose was to assess consumer interest in, and the awareness and affordability, of LED lighting. The LED Pricing Pilot addressed these two issues: price elasticity of the product and optimal price point.

The program was implemented as a store-based approach to test five price points, and both income and geographical area were key variables in the design. Stores were selected into the treatment and matched-control groups based on the median household income of the neighborhoods in which they were located. Price levels were assigned to individual stores, so that each price level was represented in all income areas. Stores with a given price point were separated geographically by other stores that had a different price point, so that customers could not travel short distances to find a lower price. This resulted in having buffer stores separate the treatment stores from the standard-pricing stores. The control stores were used to account for non-program factors other than the program incentives.

The 64 stores of the name-brand retailer selected for the trial are mapped in Figure 1. The colored dots are stores in the pricing trial. The gray dots (marked “not used” in the legend) represent the buffer stores, while the members of the control group are shown as black dots (marked with zero incentive in the legend). The colored dots indicate the stores in the pricing trial, and each color represents a different pricing trial incentive. The incentive levels start at $5 and increase in $5 increments up to a $25 incentive.

In consultation with the SCE program manager, evaluators worked closely with the program staff to set the various incentive levels, such that the price points being tested were set significantly less than the current retail price in the control group stores. The implementation of the five test price points were staggered so that not all price points were in effect at the same time.

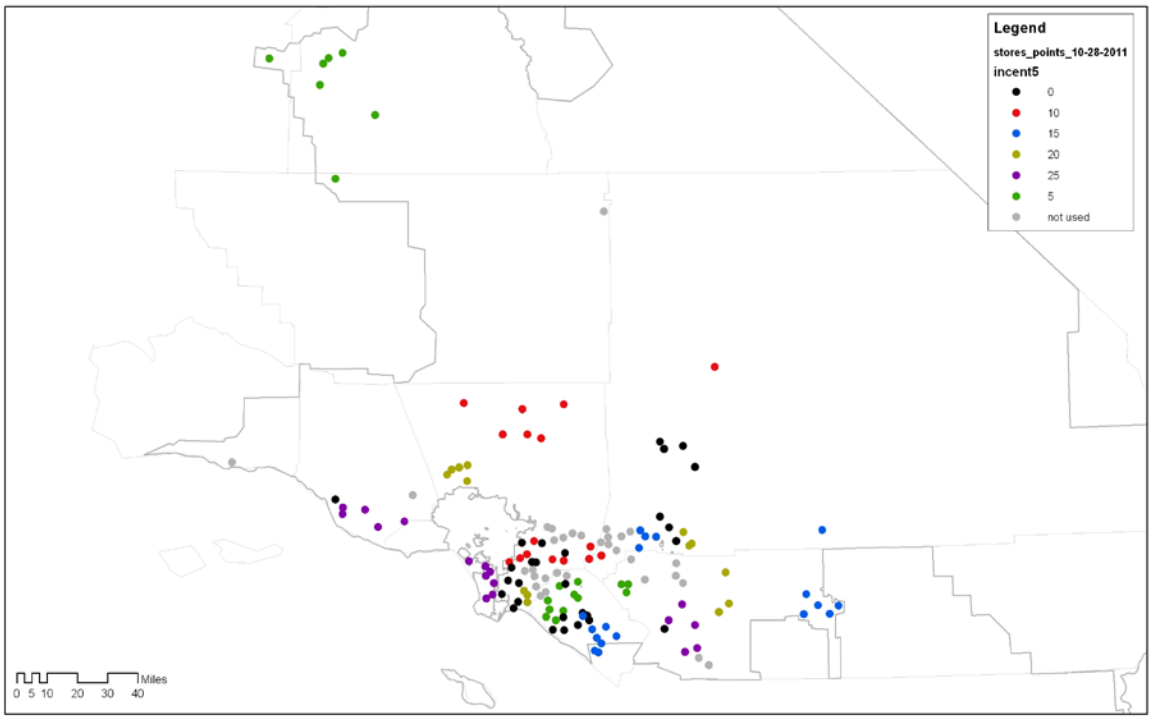

Figure 1. LED Pricing Trial Participant and Control Group Store Locations

The goal for the trial was to sell 120,000 units, using the five test price points. The primary challenges in this trial were these: (1) the differences between the planned evaluation design and the actual implemented pilot; and, (2) how best to model the sales behavior.

Differences between the planned evaluation design and evaluation of the implemented pilot. Even with a pre-launch evaluation design, there were three key challenges in the selection of the control and treatment groups, and these led to changes in the pilot implementation—and, subsequently, the evaluation:

- ***The selection of treatment stores had to be flexible*** to accommodate the retailers' ability to vary the prices across their stores.
- ***Selecting a set of store-and-incentive combinations***—based on the geographic mapping of the stores—to ensure the stores locations were sufficiently distant so that customers did not engage in cross-shopping for a lower retail price entailed an intensive, collaborative effort between the evaluators and the SCE manager.
- ***Modeling sales behavior*** required collecting key data elements for both the treatment and control groups. Participation in the program required providing specific data, and retailers could not participate unless they agreed to provide these data. (As of April 2012, data were still being collected, and modeling had not been conducted.) To have confidence in the analysis and results, it was critical to obtain appropriate baseline data from the retailers (including normalizing data such as sales rates, foot traffic, and store size) before the start of the trial.

One of the key lessons applicable to other pilots was the amount of time involved (several months) in obtaining the retailers agreement to try these different incentive levels at their stores, as retailers were required to provide sales data for the selected stores. Thus, the program design needed to be flexible enough to accommodate retailer's input in the store selection for varying prices at these store locations.

A summary describing how evaluators approached selection of the treatment and control groups using multiple stores owned by one retailer is provided in Table 1.

Table 1. Summary of Treatment and Control Group Development

Group	Number in group	Incentive Level	Key Selection Criteria	Proposed Approach	Data Requirements	To Do
Treatment 1	36	$5	Median income & geographic location	Select geographically proximate stores covering all retailers, income levels, and proximities	Need sales data, including overall floor traffic and pre-trial data	Model sales data for each group
Treatment 2	32	$10		Same as above	Same as above	
Treatment 3	26	$15		Same as above	Same as above	
Treatment 4	25	$20		Same as above	Same as above	
Treatment 5	29	$25		Same as above	Same as above	
Control	48	$0		Same as above	Same as above	

The Appliance Recycling Retailer Pilot Program. In 2010, SCE added a new Retailer Pilot delivery to its 18-year-old Appliance Recycling Program (ARP). ARP offers its customers an incentive and the free removal of their old, inefficient refrigerators and freezers. Customers contact the program to request the pick-up of qualified appliances, which are then recycled in an environmentally responsible manner and fully removed from service. Since inception, ARP has recycled over 980,900 refrigerators and freezers in SCE's service territory.

The new Retailer Pilot (which ended in September 2011) intervened in a different market from the standard ARP—it targeted customers who bought new refrigerators from a retailer. The existing refrigerators of these customers were removed from the home and recycled out of service.

The trial program and standard program operated simultaneously, and SCE hypothesized that the two types of the program would have different effects in terms of haul-away volume and cost-effectiveness. Hence, it was important to test this hypothesis using a quasi-experimental design.

Implemented through a brand-name retailer, the Retailer Pilot gave customers the opportunity to sign up for ARP when purchasing a new refrigerator. If, when the new unit was delivered, the retailer's operators determined that the old unit was program-eligible, they removed it from the household, and delivered it to a program partner who recycled it, permanently removing the unit from service. SCE then provided customers an incentive for recycling the old unit. By avoiding individual household pick-ups, the pilot was expected to lower the program implementation cost.

Before the launch of the trial, program evaluators posed the following researchable questions:

- What is the baseline condition at the retailer stores before the pilot program intervention?
- What data are needed to test the hypothesis that the ARP trial attracts more customers who would not have participated in the standard program? (That is, did ARP achieve a lift in volume from a new delivery approach?)
- How will it impact the composition of the recycled units in a way that would affect program savings and cost-effectiveness?

One complication to the Retailer Pilot's design was that the participating retailer already included a recycling effort in their standard services (specifically, a haul-away service with purchase and delivery of a new refrigerator). Conceptually, it was difficult to disaggregate the different program and non-program effects on the volume of recycling. Hence, a treatment and non-randomized matched control group program design was developed to reliably measure and to disaggregate the trial program effects from the haul-away service offered before the pilot program. Thus, it was important to design the pilot and evaluation to account for this effect.

The pilot included nine treatment stores and six comparison stores of the same name-brand retailer. Before the pilot implementation, 12 months of store data were collected for each of the 15 treatment and control stores, detailing the number of refrigerators delivered and hauled away. The same data points were then collected for the duration of the pilot (approximately 10 months).

Figure 2 shows how the pilot's design provided an opportunity to collect and analyze the data to measure the effects of the Retailer Pilot program.

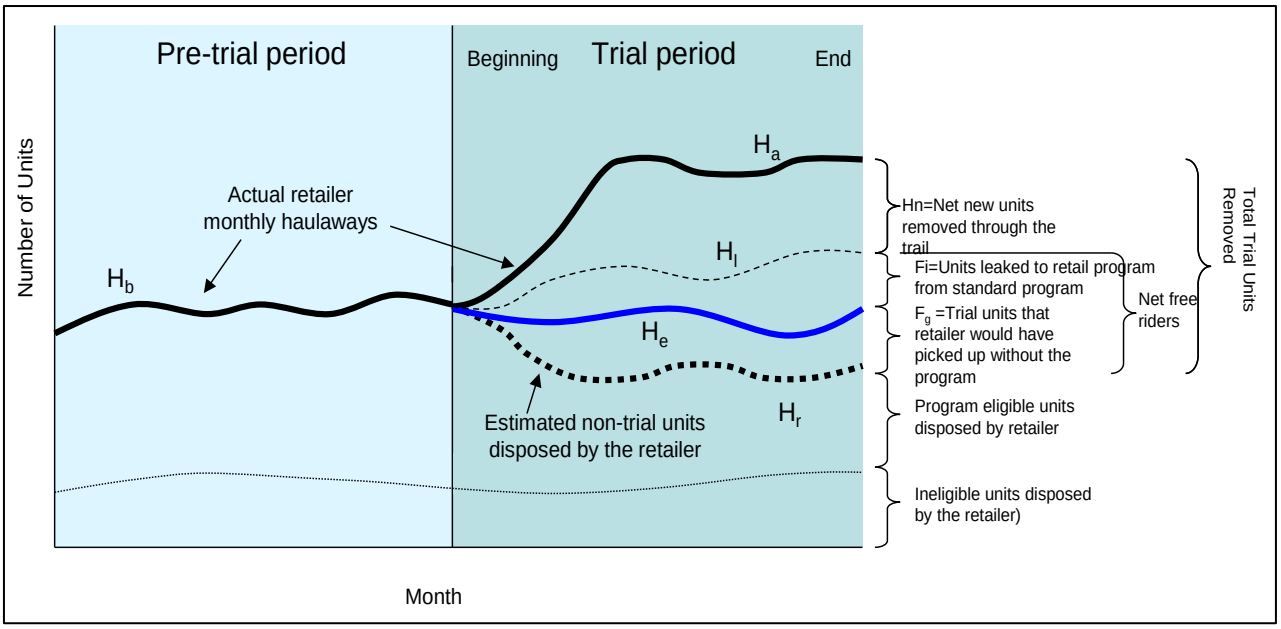

Figure 2. Effects of the Pilot as Evaluated

Showing both pre-trial and post-trial haul-away volume, Hb and He are the baseline and estimated recycling haul-aways by the retailer before and during the trial, respectively. Ha is the actual volume recorded for haul-aways by the retailer during the pilot, (including both program and non-program haul-aways). Hr measures the expected drop in the volume of the retailer haul-aways.

Fg are the trial units the retailer would have removed even if the customer did not receive the incentive, and Fi are the units that "leaked" from the standard program to the trial program. The net freeridership of the trial program is determined by adding Fi and Fg. It is the effect of the pilot (Hn) —net units or pure lift in units recycle—that is of major interest.

Shown below, Table 2 provides a summary of how evaluators approached selection of the treatment and control groups using multiple stores owned by one retailer.

Table 2. Summary of Treatment and Control Group Development Using One Retailer's Stores

Group	Number in group	Key Selection Criteria	Approach	Data requirements	Questions/issues	To Do
Treatment	9	Stores allocated to three cluster groups based on demographics: • Group 1, lower to middle class • Group 2, middle to upper-middle class • Group 3, high-income, suburban, single-family or young urban professional in multi-family	Used cluster analysis to discriminate among the retailer's stores	• Five demographic characteristics for a store's ZIP code and all adjacent ZIP codes • Retailer sales, stimulus removals, and haul-away data (Feb. 2008 to Sept. 2011) • Standard and retail ARP participation data Feb. 2008 to Sept. 2011	No store sales data or market data were available at the time of store selection to inform the selection. Hence, demographic data were used	Later analysis of actual sales data indicated that the approach was reasonable if not perfect
Control	6	Same as above	Same as above	Same as above	Same as above	Same as above

The complexity of the measurement of pure effects for the Retailer Pilot rested heavily on the pre-launch haul-away data for both the treatment and the comparison stores. However, the retailer did not always have these data in a form that was readily retrievable for all the stores (both treatment and comparison). In some cases where the data were not accessible, the evaluator performed an on-site, manual review of records to ascertain the haul-away volume for the subject store.

Without the pre-launch design evaluation considerations, it would have been extremely hard

to disaggregate the net effects of the program. Despite delays in the program start, which were related to securing access to the retailer baseline haul-away data, working out the trial design details early on was instrumental to measuring the program net effects.

Post-Launch: Behavior Change at Work Pilot

The Workplace Conservation Awareness pilot (WCA), sponsored by British Columbia Hydro (in Canada), invited some of their largest customers in their commercial and institutional sectors to participate in a program that intends to create a work environment that advances and supports energy-efficiency. The WCA assumes that helping building occupants take steps to save energy at work will coalesce into an organization-wide ethic that produces measurable and persistent energy savings. The WCA evaluation illustrates how incorporating evaluation after a pilot is underway is more likely to uncover surprising challenges that could derail the effectiveness of the evaluation and delay its results.

In 2007, during the first phase of the pilot, the WCA was offered to 10 organizations. At that time, it received a process evaluation that included interviews with participants and sponsors. The process evaluation showed strong support from participants for continuing the program and suggested the pilot was likely achieving savings. However, since an impact evaluation was not conducted, the efficacy of launching a full scale program could not be established.

In 2010, the pilot expanded—by invitation only—to large utility customers across these sectors: elementary/secondary schools, colleges/universities, office buildings, retail/hospitality, municipalities, and healthcare.

The WCA currently (April 2012) involves approximately 40 large organizations, each with multiple buildings. The participating organizations receive an annual budget of from $1,000 to $15,000 to fund program activities. The organizations also receive support from an outside consultant who advises them on program strategy and implementation. These consultants also provide regular updates to BC Hydro about each organization's activities and progress.

In 2011, after the second round of the WCA pilot was well underway, evaluators were asked to design and conduct both an impact and process evaluation. The impact evaluation is retrospective, involving a billing analysis of consumption data for participating and control group buildings (where possible) to estimate the pilot's savings impacts. The process evaluation assesses the following: (1) both past and current program delivery and performance; and (2) the impacts of planned changes to the pilot. The evaluation's objectives are to:

- Estimate the net electricity savings impacts of WCA;
- Investigate whether energy savings from WCA participants equal or exceed 5% of total controllable building consumption;
- Investigate building occupant perspectives: awareness, knowledge, attitudes, behaviors, and satisfaction related to the program; and
- Identify ways to improve WCA prior to an even broader launch.

The evaluation plan calls for a quasi-experimental approach for the impact evaluation, using a non-equivalent control group design for the impact evaluation. A control group of buildings from BC Hydro's non-participant population would be matched to participating buildings on the basis of size, sector, annual consumption, location, and age. The objective was to represent as closely as possible what the participants' consumption would have been had they not participated in the program. While recognizing that finding non-residential control group buildings can be difficult, the evaluators were hopeful the group could be established fairly simply.

For the process evaluation, the budget did not allow for a control group of nonparticipants, thus making the pilot a combination of experimental and non-experimental approaches. The process evaluation would interview key actors (known as energy champions) within each participating organization to understand the energy-saving activities of the occupants, to find out what successes

and challenges occurred, and to compare previous energy-efficiency awareness and behaviors to post-pilot awareness and behaviors. In addition, each organization was asked to conduct a pre- and post- survey of their workers. The evaluators planned to use the data to explain variations in impact evaluation results and to provide more evidence of the effectiveness of the program. The evaluators planned to produce a set of early evaluation findings to test the impact analysis scheme, pinpoint the biggest savers (if any) among participants, and connect those results to process findings.

Based upon initial discussions and data exchanges with the research sponsors, however, and work on the early findings, evaluators discovered program realities that required revisions to their approach. These discoveries affected both the process and impact evaluations.

On the process side, issues related to data quality and access to participants: (1) the pre- and post-survey data was not available or had not been collected for all participants, and where data existed, it had not been analyzed and was not ready for any standard analysis; (2) evaluators did not have these data available to help understand how former occupancy views and behaviors compared to current views and behaviors; (3) in addition, while evaluators knew about the high sensitivity of protecting consumption data, they were not aware of the need to get written permission from the Canadian respondents to be able to analyze telephone interview responses in the U.S.; this created another set of steps needed to complete the interview analyses.

The largest issues, however, emerged on the impact side: (1) evaluators and program staff did not know which workers were involved with the program in the individual buildings; (2) there was not a one-to-one correspondence between electricity meters and buildings; (3) there was an insufficient amount of billing data; (4) there was a lack of viable control group candidates; (5) the organizations did not enter the program at about the same time, as originally thought, but in fact a rolling initiation of participants occurred, so that the impact analysis was delayed by six months.

The construction of the impact evaluation control group, however, created the largest challenges, and these challenges are the focus of the rest of this section. Further discussions with program managers revealed that devising a valid control group of buildings would be more challenging than expected for several reasons.

- First, the method used to target and screen participants (i.e., large customers who were recruited to participate) introduced the possibility of two types of self-selection bias: (1) that building occupants were more motivated to save energy, and/or (2) that recruited buildings would be likely to have more opportunities for saving energy than the general population.
- In addition, for some participants, no suitable control group buildings were likely to emerge because the building types were unique, and all suitable buildings were participating in the pilot (such as ferry system buildings and hospital complexes).

Taken together, these challenges meant that the evaluators needed to adopt an approach to securing a viable group of control buildings which varied across the six organizational sections. Therefore, the evaluators applied these general principles for constructing the control groups:

1. Perform a match based on: (1) characteristics of buildings, including electricity consumption prior to the program; (2) whether the building is heated with electricity; (3) whether the building was a participant in a prior BC Hydro program; and (4) location, which may be correlated with attitudes and knowledge. If available, matching would also rely on building structure, occupancy, and energy end uses.
2. Use, where possible, non-participant buildings from the "parent organization." This will minimize the potential for the first kind of selection bias and maximize the likelihood that buildings in the treatment and control groups will have similar work cultures, energy use patterns, and interests in energy conservation.
3. When suitable non-program buildings from the parent organization are unavailable, matches outside of the parent organization will be based on buildings that have similar characteristics.
4. Obtain the appropriate control group buildings through these means:

a. Asking organizational leads (energy champions) if they know of non-participant buildings that have similar physical and operating characteristics;
b. Analyzing data about building energy use, location, and past program participation;
c. For school districts and municipalities, analyzing local demographic and economic data to identify areas with similar population characteristics.

5. Identifying at least one building for the control group for each treatment group.

Table 3 summarizes how evaluators are applying this approach. The number of current or future participants is shown in the column labeled "Number in group."

Table 3. Summary of Control Group Development

Sector	Number in group	Analysis by Subsector?	Proposed Approach	Non-participant Building Data Requirements	Questions & Issues	To Do
Advanced Education	16	Maybe: Classrooms, computer labs, faculty office buildings; gymnasiums; industrial arts facilities	Match participant buildings to similar non-participant buildings from same or different university	Participant and non-participant-university, college buildings	In billing system, can we identify a building type, e.g., gymnasium? May not have enough observations to estimate savings for some types	Obtain data for non-participant buildings
Government & Municipal	152	Yes: Libraries, community center, other	Match participant buildings to similar non-participant buildings from same or different municipality	Participant: municipality buildings and similar buildings for non-participant municipalities		Identify similar municipalities
Health Care	36	No	Match participant buildings to similar non-participant buildings from same or different health care provider	Buildings of participating health care providers or buildings of non-participant health care providers	Are there non-participating health care providers?	
Hospitality	132	Yes: ferry terminals (72), restaurants (30), ski resort building facilities (28)	Match participant buildings to similar non-participant buildings	Non-participant restaurant chains and ski resort buildings	No non-participant ferry terminals; ski resorts buildings w/unique functions; only one casino and hotel.	Identify comparable restaurant chains
Property Management	31	No	Match participant buildings to similar non-participant buildings from participating property management companies	Non-participant buildings of participating property management companies	How did property management companies select buildings?	
Schools	121	Yes: elementary, middle, secondary	Match participant buildings to similar buildings in non-participating districts	Buildings of non-participant school districts		Identify comparable school districts

This ongoing evaluation of the WCA pilot illustrated how evaluation realities can emerge that force a different approach to ensuring that results will be helpful for guiding the program to its next phases. While the evaluation approach initially assumed control group buildings would be fairly easy to locate, the program's targeting of willing large customers introduced potential sources of bias and difficulties in finding buildings that matched the treatment groups. The evaluators found they needed to expand their approach to finding control group buildings that would reduce bias, ensure accurate energy savings estimates, and allow the pilot results to transfer to a wider array of customers.

Conclusions

Pilot and trial programs often fail to inform the next steps for program development, in part due to the nature of pilots (that is, they are specialized, short term trial runs), in part due to pilot designers being faced with how to test complex program ideas on limited budgets, and in part because evaluators are not equipped to ask the right questions or do not have the right tools to evaluate pilots more effectively. There is not one single best method that should be used to conduct evaluations. In fact, all circumstances are unique, and evaluations must fit within specified criteria, but the goal is to provide useful, relevant, and actionable information.

While it is clear that under optimal conditions we should weave together pilot and evaluation design up-front, this often is not the reality. And while experimental design approaches provide a welcome framework that can make pilots more manageable and help designers and evaluators think more clearly about how to ensure pilot results are transferable, many situations do not allow for experimental design approaches. In fact, even in our examples, the construction of adequate control groups proved to be a consistent challenge.

Based upon our review of common challenges in evaluating pilot programs we suggest evaluators do the following:

1. Build an evaluability assessment into the front end of each evaluation. Using the questions listed in Table 4 as a guide, dig down as deeply as possible into the program logic, design, and logistics to make sure that mechanisms are in place to allow for a robust evaluation.
2. Include an interim round of EM&V analysis in the evaluation design. Despite conducting an evaluability assessment at the front end, pilot conditions may change or details may not be uncovered that will affect evaluation effectiveness. Conducting an early (interim) evaluation or doing a 'dry run' of the evaluation analysis is another safeguard for ensuring success. This will uncover, for example, data shortcomings, inaccurate assumptions, and unknown issues that can be resolved to enable the full evaluation to proceed.
3. Throughout the evaluation process, bear in mind how results can be usefully generalized and transferred to a larger program, since most pilots want to test the waters for a workable program despite having elements that will not be transferred. For example, if managers are considering dropping certain program elements from the next generation of the program, surveys of pilot participants might try to better understand their interest in the program if those elements were removed or changed significantly.
4. Finally, remember that the purpose of the evaluations of pilot programs is to utilize the results. The job of the evaluator is to convey accurate information and to help program designers (and other stakeholders) understand the relative importance of the results and their practical application to next steps. For some evaluators, this may mean they need to depart from their usual cautious researcher stance and become advisors and collaborators in the future.

An Appendix: Using Evaluability Assessments in Pre-Launch and Post-Launch Evaluations

Although energy efficiency program evaluation has been around for many years, our experience has found underlying issues that affect program management and program evaluation. This is true for both pilot programs and mainstreamed programs. Commonly found issues include the lack of a clearly defined problem, a misunderstanding of the market, a poorly described (or missing) program theory and logic model, inconsistent tracking, little or no documentation of baseline conditions, and missing contact information for participants, non-participants and stakeholders. One of the reasons there are data gaps is that oftentimes programs (pilots and mainstream) focus on the

production aspect of the program, e.g., measure installation or workshops or other activities, and everything else is secondary.

Data are critical to a successful evaluation, and pilots typically require more, not less, data than mature programs. Without these data we cannot properly attribute savings to the program nor determine whether it was cost effective, nor provide recommendations to ramp it up or change direction. Without essential data, evaluators cannot, for example, contact participants, select and confirm the validity of a control group, determine that the sample selected from the treatment group was representative of the population, or disaggregate activities to attribute savings to the programs.

As evaluators, we understand that program managers and implementers may not have the same perspective and might not understand the data needs of evaluators. To that end, we created the evaluability assessment (EA) tool to guide program designers, planners, implementers, and evaluators in defining the problem, defining the research objectives, and collecting the data necessary for a meaningful evaluation.

- For program managers, the EA provides a roadmap of information and data requirements that should be part of their management plan.
- For contractors and implementers, it provides an overview of responsibilities and requirements prior to program implementation.
- For evaluators, the tool provides a systematic template to review program documentation and data tracking systems at an early stage to identify gaps that may affect evaluation plans and strategies.

(A complimentary second tool for program proposers, managers, and implementers, is often used to explain the rationale for the data and information requested for program specific evaluations.)

A summary of EA key categories of questions modified for pre-launch and post-launch pilot program evaluations are shown in Table 4.

Table 4. Evaluability Assessment Table for Pre-launch and Post-launch Evaluations

QUESTIONS	PRE-LAUNCH	POST-LAUNCH
Who is the intended audience for the evaluation?	Does the evaluation audience include the same people impacted by or expected to be involved with evaluation activities? Is the intended evaluation audience involved in making decisions about focus and priority?	Has the evaluation audience been identified? Have they identified key items of interest? Is the intended evaluation audience involved in making decisions about focus and priority?
What is the research question?	Is there a problem that needs to be solved?	What is the problem the pilot is solving?
Is there an explicit program theory?	Has the program theory addressed the problem and defined what needs to be measured?	Is there a program theory that states how the pilot will solve the problem?
Is there a logic model?	Has a logic model been developed? What are the indicators of success? Can success be measured using these indicators?	Is there a logic model? Are the indicators of success defined and documented? Can success be measured with the data collected?
Is there a description of the target market?	Does the pilot correctly characterize the market and relationships between the market actors? Is contact information recorded for key market actors?	Who is the target market? Can the market actors be identified? Was contact information documented?
Is the program's intended audience defined?	Who are the intended program participants and non-participants? Are the groups defined? Is contact information available?	Is there a clear definition of program participants and non-participants? Is contact information available?
Are there key market barriers that would inhibit participation?	Are there key market barriers to participation? What are they? How will these barriers be addressed?	Are there key market barriers that affected participation? What were they? Were these barriers being addressed? What kind of problems emerged?

QUESTIONS	PRE-LAUNCH	POST-LAUNCH
Will the program be delivered with trade allies or stakeholders?	What are the roles of trade allies and stakeholders? Are the trade allies defined well enough to identify potential participant and non-participant trade allies or stakeholders if they are delivering the program?	What are the roles of trade allies and stakeholders? Are they delivering the program? If so, were they identified? Was contact information documented?
How will the data be captured? Is there a tracking database?	How will the data be captured? Is there an electronic tracking database that includes participant and non-participant contact data? Does it capture their program-related activities?	Is there a tracking database? Does it contain contact information, activities, and program measures installed? Does it record non-participant contact information and activities?
Is there a description of the operational staff?	How many staff will operate the program? Where will they be located? Are their roles clearly defined?	Who are the operational staffs? Are responsibilities understood by all who touch the program?
Is there a marketing plan?	Does the marketing plan target the intended market with the appropriate message that could elicit participation? Is there a way to measure effectiveness?	Was a marketing plan developed as an integral part of program design? Who deployed the plan? Did it result in recruiting participants? Is there a way to measure marketing effectiveness?
Are program activities and assumptions documented?	If the pilot includes installation of specific energy saving equipment, are there instructions to document locations so inspectors can find measures during on-site verification?	If the pilot includes installation of specific energy saving equipment, are specific locations documented? Can they be found to verify installation?
Are energy savings assumptions documented?	How will the energy savings be calculated? Have assumptions been documented?	Does the program document energy savings calculations? Are energy savings assumptions documented?
What type of evaluation is planned?	What evaluation protocols and methods can be or should be used? How do these fit within the structure of the pilot program?	What kind of evaluation can be conducted, given available data, and give n the data that can still be collected? How does it fit with the current program?

References

Bronfman, B., Samiullah, S., West, A., et al. *Integrating Evaluability Assessment into the Program Planning, Implementation and Evaluation Process: Case studies from Southern California Edison's IDEEA program portfolio*. Association for Energy Service Professionals. 18th National Energy Services Conference Proceedings, January 2008.

Caruth, D., Bardeaux, J. *Uses of Evaluation Findings: Taking a First-Year Industrial Program to the Next Level.* International Energy Program Evaluation Conference proceedings. June 2012.

Murray, C., Fawcett, J., *Dress Rehearsal: The Importance of Pilots to Successful Policy*. International Energy Program Evaluation Conference proceedings. June 2012.

West, A., Bronfman, B. *Magic Ingredients in Evaluability Assessments*. International Energy Program Evaluation Conference proceedings. June 2009.