Comparison Groups for Whole Building Program Impact Evaluations: They are Harder and Easier Than You Think

Dr. Katherine V. Randazzo, Opinion Dynamics, Carlsbad, CA
Dr. Richard S. Ridge, Ridge & Associates, San Rafael, CA
Seth Wayland, Opinion Dynamics, Oakland, CA
Brian A. Smith, Pacific Gas and Electric Company, San Francisco, CA

ABSTRACT

This paper addresses design issues specific to whole-building programs designed for achieving deep savings. We find that there are research design issues specific to whole-building programs that merit extensive discussion, and that care has not always been taken to think through those issues and document decisions about them. We also find that many evaluators are confused about the issues and definitions of key terms. Our paper focuses on comparison groups and the unique advantages and disadvantages of finding appropriate comparison groups for estimating gross and net savings for the whole-building type of program. Evaluators commonly choose the example of a single-measure program, such as an air conditioner rebate program, to think through what kinds of comparison groups are necessary to estimate program impacts. But the issues surrounding comparison groups are different for programs that include numerous measure types and where multiple, expensive measures are required to qualify for the program.

We consider: 1. The complexities in using future participants as a comparison group for estimating gross savings. For example, might future participants have installed some program-qualifying measures in the year prior to their participation, and would that lead to a savings estimate somewhere between gross and net? 2. Both the advantages and disadvantages of choosing a comparison group approach for estimating net savings for a program that requires multiple, expensive measures to qualify, and where there is a long list of qualifying measures. For example, when estimating net effects, is a customer who installed insulation, did air sealing, and replaced an air conditioner a good comparison customer for a participant who put on a cool roof and installed efficient windows? 3. The importance of checking the assumptions behind the use of any comparison group for estimating gross or net savings. This paper considers these issues and makes recommendations for choosing and documenting decisions about constructing comparison groups for this type of program.

Introduction

This focus of this paper is on identifying appropriate comparison groups for estimating program gross and net impacts for energy-efficiency programs in general, and for whole-building programs in particular. We have noticed several assumptions that researchers commonly make in thinking about comparison groups and the estimates they support:

1. The belief/assumption that the inclusion of a comparison group always provides an estimate of net impacts.
2. The belief/assumption that using future participants as a comparison group addresses self-selection factors, and is thus an appropriate source of comparison group members for estimating net impacts (since self-selection into programs is associated with free-ridership).
3. The belief/assumption that a future-participant comparison group always allows us to estimate gross impacts.

We take issue with each of these assumptions, and discuss the extent to which each may or may not be applicable in certain circumstances for both single-measure and whole-building programs. Any discussion of research designs and comparison groups requires a common understanding of key terms, so we address them next.

Establishing Common Understandings

Given our position that careful definitions of common concepts are particularly required for making good decisions about comparison groups in whole building programs, it seems appropriate to define three concepts here:

Gross Impact: Changes[1] in energy consumption that result directly from program-related actions taken by participants of an Energy Efficiency (EE) program, regardless of why they participated (Violette and Rathbun 2014, p.3).

It is important to observe that this definition only addresses the effect of the program-related actions (i.e., the installation of efficient measures or the adoption of efficient behaviors) on energy consumption. It does not address the issue of how many and what types of program-related actions would have occurred in the absence of the program, which is the focus of the next definition.

Net Impact: Changes in energy use that are attributable to a particular EE program. These changes may implicitly or explicitly include the effects of free ridership, spillover, and induced market effects (Violette and Rathbun 2014, p.3).

In other words, with the phrase, "attributable to a particular EE program" we want to isolate the savings that are caused by the program from those that would have occurred naturally, i.e. in the absence of the program, what would have occurred naturally is the counterfactual, which we discuss next.

Counterfactual: Social science methods texts such as Shadish, Cook and Campbell (2002) provide a good description of the counterfactual:

> In an experiment, we observe what did happen when people received a treatment. The counterfactual is knowledge of what would have happened to those same people if they simultaneously had not received the treatment. An effect is the difference between what did happen and what would have happened (p. 5).

The implication of this definition is that the counterfactual supports estimation of net effects, in the terminology of our industry. In the general social science literature, there is usually not a distinction between net and gross impacts; but expressed in our industry terms, most of what that literature is aimed at estimating is what we would call net effects. Thus, our position is that the term "counterfactual" is best reserved for conversations about net effects. Of course, in our industry, we often have to set baseline values for gross savings as well, but we think it is less confusing to refer to such baselines as gross effects baselines rather than as counterfactuals.

While the definitions of the terms discussed in this section are relatively straightforward, defining the nature and purpose of comparison groups is much more complex. We can identify at least five ways of establishing a counterfactual:

1. Experimental Design. Here, sample units (people, businesses, etc.) are assigned randomly to treatment and control conditions, the relevant behaviors are determined (including post-program usage) for each group, and the level of behaviors of the control group are subtracted from the treatment group provides net effects. In this situation, the control group acts as the counterfactual.
2. Quasi-experimental Design. This is a design that uses a non-randomly-assigned comparison group to attempt to provide the counterfactual such that it provides the basis for estimating program net effects.
3. Self-Reported Hypothetical. In this method, evaluators establish a counterfactual by asking participants directly what they would have done if there had been no program. We might term this a hypothetical counterfactual (Ridge et al. 2009; 2010).
4. Stated-Preference Discrete Choice. In this approach, respondents are presented with bundles of program attributes (products/programs) where attribute levels are randomly assigned to products/programs, or systematically assigned such that they are distributed in a balanced and

[1] This wording is a quote, and is widely cited. However, using the word "changes" misdirects the reader to think only in terms of pre- versus post-consumption, while really the comparison could be contemporaneous. A preferable word would be "differences." The same applies to the definition of net impacts.

orthogonal way. Respondents choose which, if any, of a given set of products they would choose. A simulator based on the discrete choice model (there are multiple methods of analysis for this type of dataset) allows the critical program attribute to be varied between presence and absence of the program. The uptake rate under non-program conditions forms the counterfactual.

5. Revealed-Preference Discrete Choice: In this approach, discrete choice analysis uses econometric modeling techniques to compare the measure adoption decisions of participants to those of the comparison group, representing the counterfactual, to assess the net impact of the program. This technique allows us to examine how customers make choices for energy efficient technologies and identify the key factors that influence these choices. An important advantage of this discrete choice approach is that it is based on how customers actually behave in real life situations. (See Train 1993 and Goldberg and Train 1995 for further descriptions of this and some other methods listed here).

This paper focuses on number 2 of this list.

Next, we discuss how the words we choose to describe our evaluation designs can contribute to the confusion. Some researchers have used forecasts of energy use that they refer to as “baseline projected”, a perfectly reasonable term, but go on to refer to it as the counterfactual baseline (e.g., Granderson 2015). Others have referred to the baseline projected used to estimate gross savings as representing what would have happened in the absence of the program, words that are usually used to describe the counterfactual in the traditional research design literature.[2] Our view, and at least implicitly, that of the larger social science community, is that the term “counterfactual” and the words used to describe it should only be used in connection with estimating what we would call net effects, i.e., the program effects. There can be other types of baselines that support estimates of gross effects, but the counterfactual is, by definition, the point of comparison for estimating net effects, since it is meant to represent what would have happened without the program. That is how we use the term throughout this paper.

Comparison Group for Estimating Net Savings: One way to describe the requirements of a good comparison group for estimating net savings in the traditional literature is provided by Rubin (1974). He introduces two relevant concepts: Stable Unit Value Treatment Assignment (SUTVA), and Ignorable Treatment Assignment (ITA). The principle of SUTVA is described best by Cox: “…the [potential outcome] observation on one unit should be unaffected by the particular assignment of treatments to the other units" (Cox 1958 §2.4). An example of a violation: non-participant spillover, where comparison group members may learn about energy-saving behaviors by talking to their treatment group neighbors. Violations of SUTVA will most frequently be illustrated by non-participant spillover, or free drivers in the energy efficiency field.

The principle of ITA, sometimes referred to as Ignorable Assignment Mechanism, is described by Rubin (2004): Ignorable Assignment Mechanism: The assignment of treatment or control for all units is independent of the unobserved potential outcomes (“nonignorable” means not ignorable). For every unit, it must be possible that that unit could have been assigned to either treatment or comparison group. Further, that treatment assignment is independent of the outcome, given the covariates (Rubin 2006, p.306). This is sometimes called ”unconfounded” or ”no hidden bias.”

Random assignment of customers to a treatment or control group would accomplish ITA, but accomplishing it in the absence of random assignment is challenging indeed. It implies a comparison group with matching to participants on essentially all variables relevant to the outcome variable of consumption or change in consumption, whether those variables are observable or not.

Standard research design texts (e.g. Campbell and Stanley 1963; Shadish, Cook, and Campbell 2002) point to multiple potential functions that must be performed in quasi-experimental designs, i.e. using comparison groups where control groups are not feasible. Multiple factors could affect the outcome variable, and therefore

[2] Although, as noted above, the social science literature does not usually make the distinction between net and gross effects, so the references to counterfactuals are only implicitly aiming at net effects.

will confound the effects of the treatment if not addressed adequately.[3] They point to the need to control for such influences as:

1. **History**: Events outside of the study/experiment or between repeated measures of the dependent variable may affect participants' responses to experimental procedures. In the EE field, history includes factors that change over time, such as weather and social/economic conditions.
2. **Selection**: This refers to the problem that, prior to participation, differences between groups may exist that may interact with the treatment variable and thus be "responsible" for the observed outcome. Selection biases can occur due to program targeting, or due to customers self-selecting into the program. In the EE field, this can include many factors, including anything that has an impact on energy use and that differs between treatment and comparison groups. This certainly includes, e.g., the types of attitudes and motivations that are associated with self-selection into an EE program.
3. **Maturation**: Subjects change during the course of the experiment or even between measurements. For example, young children might mature and their ability to concentrate may change as they grow up, or, a person's attitude toward global warming might change slowly over time making them more predisposed to reducing their energy use. Changes in the needs of a household over time could also be categorized as Maturation.
4. **Statistical Regression to the Mean**: This type of bias occurs when some subjects have extreme scores (one far away from the average) such as high energy use. For example, when customers whose annual energy use is greater than 12,000 kWh are targeted for an energy audit, reductions in energy use after participation might be due to regression toward the mean and not the program's effectiveness.
5. **Testing**: Repeatedly measuring participants may lead to bias. Participants may remember the correct answers or may be conditioned to know that they are being tested. Repeatedly taking (the same or similar) intelligence tests usually leads to score gains, but instead of concluding that the underlying skills have changed for good, this threat to Internal Validity provides good rival hypotheses. This is unlikely to be a factor in most EE programs as the participants are generally not conscious of the data gathered by evaluators to show their responses to the intervention. One exception to this can occur in our industry when surveys are used at multiple times during the evaluation.
6. **Instrumentation**: The instrument used during the testing process can change the experiment. This also refers to observers being more concentrated or primed. If any instrumentation changes occur, the internal validity of the main conclusion is affected, as alternative explanations for apparent gains are readily available. This factor, as well, is not usually an important factor in determining internal validity for evaluations of EE programs, although this statement is subject to the same exception as noted with Testing.

These influences are characterized as threats to internal validity defined as the approximate truth about inferences regarding cause-effect or causal relationships. Next, we will address how these factors are dealt with in the energy-efficiency program evaluation industry generally, and then specifically for gross and net impacts.

Design & Analytic Issues in the Energy Efficiency Field

A translation of the potential confounding influences listed above into the factors that our industry recognizes as essential when conducting a consumption analysis, could look like this list:

1. Economic & political events & trends (History)
2. Weather (History & Selection)
3. Building characteristics (Selection)

[3] When one variable is confounded with another, it means that it is impossible to separate the effects of one from the other. For example, if everyone in the treatment group is over 50, and in the comparison group they are under 50, the treatment effect will

4. Occupancy characteristics (Selection)
5. Geographic areas (Selection)
6. Motivations, attitudes, and behavior (Selection)
7. Changes in motivations, attitudes, and behavior over time, apart from program-related changes (Maturation, Statistical Regression to the Mean)
8. Naturally-occurring relevant installations & motivations to install (History & Selection, and perhaps Maturation)

An important aspect of Selection, as an influence, is that it can be thought of as being of at least two types: 1. program implementer selection (by design or by accident) and 2. self-selection into a program by participants themselves.

A model successfully controlling for Factors 1 through 8 above, whatever the design, would produce net impacts. Factor 8 is quite specific to net impacts, while controlling for Factors 1 through 7 only, will generally lead to what we define as gross impacts. A simple pre/post regression with participants only will usually control for the first six factors and provide the basis for estimating gross effects without obvious bias. The exception to this is a situation where participants experienced changes in motivations, attitudes, etc. coincident with participation in the program and apart from the program effects on those factors. If changes of this kind occur, only a series of measurements over the studied period would allow the changes to be controlled for, and this is almost always impractical. So, Factor 7 muddies the waters a bit and represents a slight weakness in the pre-post design. Further, if the program was responsible for any attitude changes that occurred between the pre and the post period, the effects of those changes on usage would be attributed to the program's gross impact. However, for most programs, it is unlikely that they would have "moved the needle" on attitudes such that they would compromise the interpretation of pre- to post-program usage change to gross effects. The gross impacts would be captured by the coefficient representing the presence or absence of program-promoted equipment, unless confounded by changes in attitudes and the like. How the equipment installation is represented in the model is a subject for another paper. Representing/controlling for Factors 7 and 8, mainly falling into the influence categories of History and Selection, when adequately controlled, would yield net effects. Note that some comparison groups can serve to control for factors 1 through 7, and the result would be gross impacts, which is counterintuitive to many people who think of comparison groups as always producing net impacts. Next, we turn to more detailed consideration of net impacts.

Estimating Net Impacts

Our position is that representing the counterfactual has different requirements or operationalizations depending on the program type. Thus, we describe how we view this for single-measure and then multiple-measure programs.

The Counterfactual—Single Measure Programs

All the factors that must be controlled to produce a valid estimate of gross program savings are equally applicable to producing net savings. Designs producing net savings are distinguishable by the need to control for additional factors.

As we have asserted before in this paper, the counterfactual is applicable only to estimating net impacts. In fact, it is central to the endeavor. In the broader world of evaluation research, it is defined as what would have happened had an investment (in a program or intervention) not been made. In our industry, this translates to what a member of the eligible population/customers would have done if the program under evaluation were not there. Thus, we either need to measure directly the hypothetical situation of what participants would have done absent the program, or we must identify a comparison group that can reasonably represent the counterfactual. As this paper is about comparison groups, we discuss that alternative in some detail, next.

What was or might have been installed. For a comparison group to support estimating net effects, it must represent the counterfactual. But measuring this concept is extremely complex. We can think of two aspects of the counterfactual: 1. motivation and 2. the action taken, specifically what was installed, if anything. So, in addition to trying to represent in a comparison group, the why of participant installations (program-influenced or not), we also have to consider *what* was installed in the comparison group, and whether that technology is a suitable point of comparison for what participants installed. Even when considering only what equipment installations the counterfactual should represent, it is complex. In some programs, there is no alternative to the program equipment, or no non-efficient alternative. In others, there are various efficiency-rated alternatives.

Specifically, in cases of equipment such as air conditioners and furnaces, the customer could replace the existing equipment with something less efficient than what the code requires, or he could purchase code-compliant equipment, or he could choose a version that goes beyond what the code requires, with or without the program's influence, and to a greater or lesser degree. Other types of program-promoted equipment are either present or absent and do not consume energy. Examples of this type are duct sealing, wall insulation, or thermostats. So, the counterfactual question becomes: What might the participating customer have installed without the program, if anything? Table 1 provides some examples of the kinds of equipment that efficiency programs might promote and some possible installations that *could* represent the counterfactual *in terms of the type of equipment installed*. As a reminder, we are talking only about single-measure programs at this point.

If we could assume that any customer who took an action in the second column of Table 1 represents what participants would have done in the absence of the program that promotes the measures in column 1, we could find a comparison group that represents the installed-equipment aspect of the counterfactual, and thus be able to estimate program net effects, provided we had also controlled statistically for the first seven factors listed earlier. This would require finding customers for the comparison group who had installed these things or, something analogous to it (efficient or not), or had the opportunity to. Finding such customers can be expensive, but not impossible, as evaluators have done this many times. However, it isn't always entirely clear what the right

Table 1. New equipment installed during program evaluated year by participants and potential comparison group members, in terms of equipment only, and excluding self-selection factors

1 Participant-Installed Program Measure	2 Installed Measure that Would Qualify a Non-Participant as a Comparison Group Member for Estimating Net Effects
SEER 17 Air Conditioner	Any Air Conditioner
SEER 19 Air Conditioner	Any Air Conditioner
Tankless Water Heater	Any Water Heater
Duct Sealing	Any house with a working HVAC system that uses ducts that were not insulated during the evaluated period?
R30 Wall Insulation	Any Wall Insulation? No Wall Insulation?
Envelope Sealing	Any house with a working HVAC system that has not been sealed?

actions would be to constitute a good comparison group customer. What is the right comparison group member for a program that promotes duct testing and sealing? Or envelope sealing? Perhaps it is the customer who has not done that work but would benefit from it? The answer isn't apparent, but it is essential to address this issue in designing a comparison group.

Awareness, motivation, and the size of the eligible population. Another central issue in finding an appropriate comparison group that will represent the counterfactual in estimating net impacts can be described as the motivation and awareness of the customer making the equipment choice; in fact, this is key. The customer who is motivated to install a program-qualified measure regardless of incentive, if aware of the incentive, is highly unlikely to refuse it. (The most altruistic, committed environmentalist might do that so that the incentive could be used to motivate a less motivated installer.) Thus, environmentally-motivated customers would naturally be under-represented in a non-participant comparison group for such a program, unless the customer was unaware of the program. But an aware customer might also refuse the rebate because they perceive applying for the rebate to be a hassle. So, a comparison group pool of customers might not be completely devoid of environmental or convenience motivated customers.

The foregoing means that the only possible comparison group member for a program would be the customer who is unaware of the program or who is aware but, for whatever reasons, chose not to participate. Over time it might become more and more difficult to find such customers. If we do find them, we have to ask if those customers have the same rate of naturally choosing efficient alternatives that the participants have. Maybe the unaware customers all live in very rural areas. Would they have the same naturally-occurring rate of choosing efficient options? Do they have the same opportunity to purchase the efficient options? The answer to both is probably, No. A design that included customers that had a different set of opportunities and motivations to choose efficient equipment compared to the participant group would fail to comply with the principle of ITA. Thus, good comparison groups are unlikely to be available unless the program meets one or more of the following conditions:

1. it was relatively new,
2. it was driven by relatively few participating contractors,
3. it is only offered in a few areas,
4. the eligible population was large, and/or
5. the level of program awareness was low.

Some programs will meet one or more of these conditions and some will not. And even if they do, the evaluator must address additional complexities. In any case, where there is a large pool of non-participants under these conditions, it becomes potentially feasible to find an appropriate comparison group by further matching and/or screening, *in terms of observable variables*. Of course, a core issue in estimating net effects is the largely unobservable factors involved in self-selection. That is the specific topic of Agnew, et al. (2017), which makes it an ideal companion paper to this one.

The Counterfactual—Whole Building Programs

The issues in representing the counterfactual with a comparison group are compounded for whole-building programs. We find that discussions of comparison groups and counterfactuals are often carried out with the example of an air conditioner rebate program, and treated as if this represents all of the various program types. We find that there are issues unique to each program type, and that it is important to consider this specifically when deciding the right approach to estimating net program impacts, including whole-building programs. In the whole building scenario, Table 1 still applies, but we have to think about the entire list of measures and how the group of measures installed under the program would be represented in the comparison group, if we wanted the comparison group to support estimating net effects. At first glance, it would seem that the comparison group measure categories (column 2, in Table 1) would have to be represented in the same proportion as their counterparts in column 1. But this is called into question when we consider customers who took some, but not all of the program-promoted measures. Is the customer who did some envelope sealing and some insulation a counterpart to the program participant who did those things plus several others all under the program? Is the customer who did the envelope sealing and some insulation a good counterfactual match for the participant who had done the same things before participating, and installed a new heat pump, and a tankless water heater, and did duct sealing all under the program?

While the issue of what mix of measures constitutes good candidates for eligible customers to be counterfactual representatives is particularly complex for whole-building program evaluators, there are some issues that make it easier for evaluating this type of program compared to single-measure programs:

1. The eligible population is likely large, this type of program is relatively new, and, because they tend to be contractor driven, there will be many customers who are not aware of the program. To participate in the program, one generally needs a contractor who is approved by the program, and there are a limited number of approved contractors.
2. Some, though not all, of the participants would be recruited into the program by a contractor who is using the program as a sales tool. Customers consulting with a non-participating contractor will not be exposed to the program and thus may be unaware of it. Because there are many contractors who are not associated with the program, there may be many customers interested in an energy-related (not necessarily energy-efficient) upgrade or renovation who are not aware of the program.
3. One could make the argument that any home upgrade or renovation is an opportunity to include energy-efficiency measures in it. To the extent that customers decide to do that outside of the program, this would approximate the naturally-occurring rate[4] of such measures in this context. To the extent that they decide not to, or never think of it at all, this would represent the other part of the counterfactual. Thus, any home upgrade or renovation could be a legitimate comparison group member for a whole-house program that would yield net impacts as long as self-selection factors are accounted for.

These three situations make the task of finding appropriate counterfactual representatives considerably easier, in the sense of finding potential members, and the possibility should not be dismissed lightly. It also implies a more complex set of decisions about which potential members should be included and excluded. The large pool of unaware renovators does not automatically constitute a net-effects-supporting comparison group. More matching and/or screening, at least, would be required to produce such a group. This is not to say that a perfect comparison group can be found. It will always be imperfect, but that is true of any approach short of an experimental design. But we think it is important not to make the perfect the enemy of the good. All the decisions resulting from these complexities, and their rationales should be documented.

Future Participants – Cohort Design - as Comparison Groups

To estimate gross savings for any program, including, and maybe especially whole building programs, it is very appealing to use future participants in the evaluated program as the source of a comparison group for the evaluated participants. This approach is sometimes called a cohort design. We describe here the structure of this approach, and then consider the pros and cons of using it.

Figure 1 illustrates the cohort design, which we have simplified to conserve space. Program Cycle 1 covers 9 months with all subjects being treated in month 5 only. Program Cycle 2 also covers 9 months with all subjects being treated in the 5th month (shown as month 14) only. For both current and future participants, we have ongoing monthly measurements covering both cycles. The measurements for program months 1 through 9, for the future participants in Program Cycle 2, serve as the comparison for months 1 through 9 of the participants in Program Cycle 1.

[4] Except for the factor of self-selection into the program, such that some customers may seek out a program because they are planning a project.

	Program Cycle 1									Program Cycle 2								
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Current Participants	O	O	O	O	**X**	O	O	O	O	O	O	O	O	O	O	O	O	O
Future Participants	O	O	O	O	O	O	O	O	O	O	O	O	O	**X**	O	O	O	O

O=Recorded monthly measurement

X=Program participation

Figure 1. Illustrating the Cohort Design

This design allows us to control for exogenous factors e.g., changes in economic and political events (history) and changes in motivations, attitudes and behavior (maturation) in the general population, both of which might over time affect energy use, and for self-selection. The latter factor is particularly difficult to account for in non-randomly assigned comparison groups. Under the right conditions, this provides the basis for a good estimate of gross impacts. This design will be effective in controlling for such exogenous factors as long as these conditions hold:

1. The program design (e.g., the mix of measures promoted, the size of the rebates, etc.) remains stable,
2. The program delivery (the mix of participating contractors, their qualifications, and training, the types of customers targeted and successfully recruited, the marketing materials and channels used) have remained stable over the period of participation for both evaluated and future cohorts,
3. Future participants have not installed any program-qualified measures in the year prior to their own participation (prior to month 14 in our above example). That is, any changes to their consumption are due only to these exogenous factors, and
4. There is sufficient statistical power (as with all comparison groups).

If the comparison group members installed any program-qualified measures, or did any work that reduced their energy use (including installing non-program-qualified equipment), the resulting estimate would move toward net impact. That is, they would to some extent represent the potential free riders among the eligible population.

Targeting the same types of customers over time increases the chances that future participants will be very similar to the evaluated participants with respect to their demographics, attitudes, energy use, and building type etc. This is important since selection factors (program- and self-) and their correlates can be an important aspect in how comparable the two groups are, and future participants can be very helpful in allowing self-selection factors to be adequately controlled. This occurs because both current and future participants will have self-selected into the program, just at different points in time.

We note here that some have found the use of a comparison group composed of future participants confusing and we find that the assumptions regarding its use rarely tested. Customer targeting can change dramatically from one year to the next. For instance, one whole-house program, had targeted coastal customers during the initial program roll-out. The evaluators recommended targeting more inland areas where winters are colder and summers are hotter, thus producing more savings for participants. This is a case where using future participants was not appropriate for estimating gross impacts.

Sometimes it is not as obvious that future participants will not provide an adequate comparison for accurately estimating gross savings. This can happen when customers make some upgrades in the year prior to participation. It is highly unlikely that the customer would have installed a full complement of home upgrades in the year prior, but they may have done *some* upgrades so that they do not meet the third condition above. The only way the evaluator is likely to discover this is if she surveys those participants, or a sample of them, to determine whether such installations are sufficient to question the accuracy of the resulting gross savings estimate.

We consider it essential that evaluators who plan to use the following year's participants as a comparison group to control for exogenous factors, test the comparability of the two groups before proceeding with the plan. Taking this recommendation seriously means that the evaluator and the PA must be flexible enough to change designs if assumptions of the planned design are not met, and could mean surveying a sample of future participants to assess their comparability. Flexibility is required because such a test could not be done until the "future" cohort has been identified, which will likely be a year after the evaluated cohort was identified. Thus, if the two cohorts are not very similar, this design becomes unfeasible and a back-up plan will be needed.

We conclude with a discussion of one other source of confusion. Traditional research design literature presents the cohort design in the context of estimating (implicitly) the net impacts of a program. As mentioned earlier, the main advantage of this design is that selection biases, introduced by adding a comparison group, are reasonably well controlled assuming *the composition of the current and future participants is similar and that the design and implementation of the intervention has not substantially changed over time.*

An example used by Campbell and Stanley (1963) is an officer and pilot training program, whereby participants in year one (cycle 1) are compared to participants in year two (cycle 2). The assumption is that training to be an officer is only available through the Army's training program. In other words, during the first program cycle, the future participants in the second program cycle (or any soldier eligible to participate in the second program cycle) could not have been exposed to any training that would have prepared them to be an officer or pilot. As a result, they provide an unbiased estimate of what members of the eligible population would have done in the absence of the program, i.e., the net impact of the program.

Thus, a traditional research design text would consider this design to produce the net effects of the officer and pilot training program rather than gross effects (if, indeed, they made that distinction). An illustrative example: the evaluation of the California 2005 Low Income Energy Efficiency (LIEE) Program relied on a cohort design as one way of estimating net savings. This design was appropriate since the future participants represented what the *larger eligible population* of low-income households would have done absent the program, which is essentially nothing since they were very likely unable to afford purchasing new equipment.[5] Evaluators who use the cohort design to control for exogenous factors in estimating gross savings should clearly explain that historically such a design has been used to estimate net impacts but it is being used, in this particular instance, to estimate gross savings, assuming that the conditions mentioned earlier have been met.

This paper has discussed designing studies to estimate gross and net savings for whole-building and other kinds of programs, attending only to observable characteristics of sample members, treatment or comparison. We address the role of self-selection only in those terms. However, a companion paper on this panel (Agnew, Goldberg, Fowlie, Train, and Smith 2017) complements this discussion by suggesting new ways to address self-selection (and therefore free ridership), especially as this might be represented by unobserved factors.

Some may read this paper and conclude that we are saying that forming valid comparison groups can be nearly impossible for estimating net program effects, whether for single-measure or whole-building programs. We actually conclude the opposite: they are possible, but assumptions and conditions really need to be tested and the issues identified and documented. And we say again, the perfect should not be the enemy of the good. Jurisdictions that demand a level of accuracy that is beyond what any evaluator can typically provide will always be disappointed.

[5] Some may argue that future participants may not represent the larger population because those who urgently need to save on energy bills would be more likely to join the program than others. Our position is that in certain programs such as low-income programs that are contractor driven and are implemented by neighborhood, with the contractor requesting permission to do the work free of charge, there will be very few low-income residents who actively seek to enter the program.

References

Agnew, K. and M. Goldberg 2009. "Getting to the Right Delta: Adjustment and Decomposition of Billing Analysis Results." *International Energy Program Evaluation Conference*.

Agnew, K.; M. Goldberg, M. Fowlie, K. Train, and B. Smith 2017. "Not just another pretty formula: Practical methods for mitigating self-selection bias in billing analysis regressions." *IEPEC Meeting*, Baltimore, MD, August 8-10.

Campbell, D.T. and J.C. Stanley 1963. *Experimental and Quasi-Experimental Designs for Research.* Chicago, Il: Rand McNally College Publishing Company.

Cook, T.; M. Scriven, C.L. Coryn, S.D.H. Evergreen 2010. "Contemporary thinking about causation in evaluation." *American Journal of Evaluation*: 31:105. http://aje.sagepub.com/content/31/1/105.

Cox, David R. 1958. *Planning of Experiments*. New York: John Wiley.

Evergreen Economics 2013. "Needs Assessment for the Energy Savings Assistance and the California Alternate Rates for Energy Programs: Volume 1: Summary Report." Prepared for: Southern California Edison, Pacific Gas and Electric, Southern California Gas, San Diego Gas and Electric and the California Public Utilities Commission.

Goldberg, M. and K. Train 1995. "Net Savings Estimation: An Analysis of Regression and Discrete Choice Approaches." Submitted to California Demand Side Management Advisory Committee, Subcommittee on Base Efficiency.

Granderson, J. 2015. "Accuracy of Existing Use Baselines, AKA Normalized Metered Energy Use." A presentation on Implementation of AB802, California Public Utilities Commission, January 26-27, 2016.

Huitema, B. E. 2011. *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-experiments, and Single-Case Studies*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Illinois Commerce Commission 2016. Illinois Statewide Technical Reference Manual for Energy Efficiency Version 5.0: Volume 4: Cross-Cutting Measures and Attachments: Attachment A: IL-NTG Methodologies. Prepared for the Illinois Commerce Commission by the Illinois Evaluation Teams (ADM Associates, Cadmus Group, Itron).

Kennedy, P. 2008. *A Guide to Econometrics*. Malden, MA: Blackwell Publishing.

Mohr, L. B. 1995. *Impact Analysis for Program Evaluation*. Thousand Oaks, CA: SAGE Publications.

Ridge, R., K. Keating, L. Megdal, and N. Hall 2007. "Guidelines for Estimating Net-To-Gross Ratios Using the Self Report Approach." Prepared for the California Public Utilities Commission.

Ridge, R., N. Hall, R. Prahl, G. Peach, and P. Horowitz 2013. "Guidelines for estimating net-to-gross ratios using the self-report approach." Prepared for the New York Department of Public Service.

Rossi, P. H., M. W. Lipsey, and H.E. Freeman 2004. *Evaluation: A Systematic Approach*. Thousand Oaks, CA: Sage Publications.

Rubin, D. 2004. “Basic Concepts of Statistical Inference for Causal Effects in Experiments and Observational Studies.” Department of Statistics, Harvard University.

Rubin, D. “Estimating causal effects of treatments in randomized and nonrandomized studies.” *Journal of Educational Psychology, Vol. 56, 5*, 1974: 488-701.

Rubin, D. 2006. *Matched Sampling for Causal Effects*. Cambridge, MA: Cambridge University Press.

Shadish, W. R., T. D. Cook, and D. T. Campbell 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin Company.

Train, K. 1993. “Estimation of Net Savings from Energy Efficiency Programs.” Submitted to Southern California Edison Company.

Train, K., M. Goldberg and K. Agnew 2017. “Compared to What? Practical Tools for Consumption Data Analysis Mitigating Self-Selection Bias.” Submitted to Pacific Gas and Electric Company.

Violette, D. and Rathbun, P. 2014. “Estimating Net Savings: Common Practices,” Chapter 23 of the Uniform Methods Project, National Renewable Energy Laboratory.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.