Practical Guidance for Selecting Opt-In Research Designs: Addressing Methodological Trade-offs and Avoiding Common Pitfalls

Lucy Morris and Brian Arthur Smith, Pacific Gas and Electric Company, San Francisco, CA

ABSTRACT

The recent growth of opt-in behavior-related demand response (DR) programs has not been mirrored in the energy efficiency (EE) domain. The evaluation challenges of opt-in EE behavior programs have resulted in their infrequent use. Effects of opt-in behavior DR treatments (e.g. optional time-of-use rates) are typically large enough to be identifiable with a program design with moderate statistical power. By contrast, opt-in EE behavior programs typically yield small effect sizes, and thus have a high risk of failing to identify a legitimate effect (type II error) unless they are designed to have high statistical power.

As interest rises in opt-in, behavior-based EE programs, program managers are being called upon to run trials in the field of behavioral science, often without having a good understanding of experimental design. A high-level understanding of the relative strengths, weaknesses, and underlying assumptions of experimental and quasi-experimental approaches is critical, but the technical reference materials available are targeted at the evaluation community and econometricians.

This paper leverages the authors' "in-the-field" experience providing evaluation support and guidance to utility EE program managers running a number of opt-in behavior programs and trials. Its goal is to provide guidance on the key research questions, to assist practitioners in selecting the research design most appropriate to address their business needs and their corresponding methodological tradeoffs. Specifically, the paper provides an overview of the three most commonly-used research methodologies: Randomized Control Trials (RCTs), REDs, and Quasi-Experimental Designs, and some practical issues to be considered for each.

Methodological Context

Many of the current statistical techniques used to assess the likelihood of a set of observations occurring in an experiment if the null hypothesis is true (that there is no relationship between two measured phenomena) are based on the seminal work of R.A. Fisher (1935). The work of Campbell and Stanley (1963) on experimental and quasi-experimental design remains a relevant resource even 50 years after its publication. As Campbell and Stanley observe:

"*Good experimental design is separable from the use of statistical tests of significance. It is the art of achieving interpretable comparisons…if the comparison is interpretable, then statistical tests of significance come in for the decision as to whether or not the obtained difference rises above the fluctuations to be expected in cases of no true difference for samples of that size. Use of significance tests presumes but does not prove or supply the comparability of the comparison groups or the interpretablity of the difference found.*"

This paper focuses on considerations around program design to support interpretable comparisons. The complex detail of the various statistical analytical approaches and methods then used to judge the significance of any differences found is outside the scope of this effort.

Essential to the valid interpretation of comparisons is acknowledgment of the influencing factors that have (and have not) been controlled in any given study. Because of their ability to control for so many of the factors that could influence or bias the findings, Randomized Control Trials (RCTs) are often referred to as the "gold standard" of research design. An RCT's "*distinguishing feature is…that*

the various treatments being contrasted (including no treatment at all) are assigned to experimental units by chance…If implemented correctly, random assignment creates two or more groups of units that are probabilistically similar to each other on the average. Hence, any outcome differences that are observed between those groups at the end of the study are likely to be due to treatment, not to differences between the groups that already existed at the start of the study." (Shadish et al. 2002, 13).

A relatively recent variant of experimental design being used to evaluate program effectiveness is a Randomized Encouragement Design (RED). An RED is an RCT that involves offering encouragement to those assigned to a treatment condition. A RED requires that random assignment occur prior to any solicitation or encouragement provided to those assigned to a treatment condition (Cappers et al. 2013.) Those assigned to a treatment group are encouraged to opt-in, and the RED rests on the fundamental assumption that members of groups are encouraged to participate will have higher rates of opting in than those persons assigned to the control group. The result is a positive correlation between being in the encouraged group and being treated, which then can be used to estimate the treatment effect on customers who were in fact treated.

Of course there are many real-world settings where a researcher is unable to control relevant factors sufficiently to be able to adopt an experimental design that requires random assignment. In cases where experimental designs are not feasible, a quasi-experimental design can be the best alternative. A quasi-experimental design is one in which treatment and comparison groups are generated in a non-random manner. When the comparison group can be created to resemble the treatment group on factors relevant to program outcomes, reliable inferences can be made about program effects. Quasi-experimental designs are considered less desirable than RCTs because of the practical challenges involved in creating two (or more) groups that are equivalent on key characteristics relevant to the program outcomes (besides exposure to the treatment). Where key characteristics cannot be matched or controlled statistically, the comparisons risk being biased in their estimate of program effects.

Utility Program Context

Recently there has been rise in Program Administrators (PAs) designing EE programs whose effectiveness is reliant on triggering complex consumer behaviors—well beyond the straightforward behaviors associated with purchasing and installing widgets. For example, if a residential customer purchases and installs a "smart" thermostat in exchange for an EE program incentive, what type of thermostats did she replace, how does her family members use the new thermostat differently from the replaced unit, to what degree will she enable the new thermostats' "smarts" and how long will the behavioral changes endure? Behavior-based EE programs are challenging because of the high variability of human behavior and the complexity of behavior-dependent technologies and systems. Moreover, the necessity for utilities to avoid "double-counting" of impacts results in the requirement that evaluations tease apart savings resulting from equipment/widget-related behaviors (that might have been attributed to existing EE measures) from "other behaviors" (triggered by a behavior-based intervention). It is the small effect sizes (that is, savings ranging from 1% to 3% of total electricity and/or gas usage) that are typical of these "other behaviors" that pose such major evaluation challenges. In the face of small effect sizes, utility evaluators often turn to RCTs because large sample sizes provide high statistical power that maximizes the chance of reliably identifying small effects and avoiding Type II error (failing to reject the null hypothesis that there is no difference in energy use between treated and control groups when, in fact, there is a treatment effect).

However, the reality is that utility evaluations face a number of constraints which influence EE trial and program design; budget restrictions and customer experience concerns top the list. The budget constraints usually mean that PAs cannot follow the recommended best practice of conducting randomized experiments with multiple treatment degrees and types, and their interactions, and instead

must often take the less nuanced approach of comparing the effect of one treatment against no treatment. A recent exception to this is PG&E's Home Energy Reports (HER) program, which is designed as a large opt-out RCT and has conducted various trials using factorial design. For example, a trial was recently conducted whereby a random selection of HER recipients received the reports only by mail (the traditional approach), while some received it both by mail and electronically. The concerns around customer experience center on not wanting to force or deny treatment to customers. These concerns often result in the adoption of a quasi-experimental approach whereby the treatment analysis is conducted with a generated comparison group rather than a randomly-selected control group.

Unique Challenges of Evaluation Design for Utility EE Opt-in Programs

Opt-in designs are those in which the customer chooses (or "self-selects") to participate. This self-selection might involve signing up to a specific utility rate program, receiving a utility rebate in exchange for recycling an old refrigerator, taking an online home audit, or purchasing a rebated product. While many PA programs are downstream rebates and thus based on the customer's purchase decision (opt-in), to-date many behavior-based programs use opt-out designs due to measurement and evaluation challenges associated with small effect sizes. For example, PG&E's HER program has, on average, a million residential customers receiving treatment, and approximately 700,000 customers in the control groups. Because the program uses an opt-out design, the utility can achieve these large sample sizes required to reliably evaluate a small effect size – between 1% and 2% of household energy use.

Opt-in, behavior-based programs pose an evaluation challenge that distinguishes them from widget-based programs (typically opt-in because the customer elects to buy a rebated widget/appliance), and from behavior-based opt-out programs (like HERs) because the savings from widget and opt-out programs are significantly easier to estimate. While widget-based programs typically rely on a specific deemed savings value associated with the concreate action of buying and installing a widget, and opt-out programs have the ability to achieve sample sizes in the hundreds of thousands, opt-in behavior programs have neither of these advantages. Behavior-based, opt-in programs face the double evaluation challenge of correctly identifying a small effect size while being unable to achieve the very large sample sizes of an opt-out program. There is the additional challenge of risking low external validity if evaluations require somehow assigning or encouraging opt-ins that might not happen if the program was operating in the market at scale. Thus, opt-in behavior programs require statistically powerful evaluation techniques to accommodate their typically small effect sizes, while not necessarily being compatible with the most powerful experimental design (i.e. RCT.)

The key challenge in estimating the impact of an opt-in program is to identify a control group against which to reliably compare energy use to the group who opts-in. Self-selection bias must be considered if customers who choose to opt-in are unique at any given point of time. For example, those who opt-in during 2014 may different from those who did not opt-in during 2014 (possibly due to attitudinal differences in technology adoption or differences in price sensitivity), but may also be different from those who opt-in during 2015. These differences may be due to situational (for example, exposure to promotions or news events) and/or dispositional (for example, personality characteristics driving a predisposition to opt-in or not) circumstances. Three common options for addressing this risk of self-selection bias are noted below. Each of these approaches involves benefits and challenges which must be weighed by the PA when designing the program or pilot to ensure the program design best supports evaluation:

- The most obvious solution to addressing the risk of self-selection bias is to create a control group from the same group of people (i.e. those who opted-in during the same time period) but for whom treatment was denied or delayed (known as recruit-and-deny and recruit-and-delay methods).

- Another option is to use a variation on RCT called Randomized Encouragement Design (RED) which avoids any denial or delay of treatment to those who want it, and instead moves the randomization element earlier in the design so that it precedes a consumer's decision to opt-in.
- A third option is to use a quasi-experimental method by identifying customers as similar as possible to the treatment group on key variables (e.g., geography, home type/size, family size, energy usage amount and patterns, etc.) and take the risk of not being able to match on any unknown "opt-in" variable that may have driven the consent to participate.

Option 1: Randomized Control Trials

Although the RCT design is strongly preferred for evaluating opt-out behavioral programs, applying the RCT design to opt-in programs can be problematic given the typical small effect sizes and the resulting need for large treatment and control groups. Relevant issues when determining whether an RCT approach is appropriate to opt-in designs include: the Program Administrator's (PA's) appetite for a full-scale territory-wide program, since an RCT would make that impossible; whether the utility is comfortable with the potential for a poor customer experience due to the use of recruit-and-delay or recruit-and-deny approaches that an RCT would require; and whether there are utility requirements that the program be made available to all customers.

How an RCT works

An RCT is designed such that each member (person, household, business, building, etc.) of the population of interest has an equal chance to be assigned to receive a treatment or not. In this design, members of the control group are denied any treatment, even if wanted. This requirement to deny treatment is what makes this design so challenging: it can be a negative experience for someone to opt-in only to be told that they cannot receive the treatment. Moreover, a member assigned to the control group may elect to source the treatment independently of the experiment.

One alternative to denial of treatment to members assigned to the control condition is to *delay* the treatment for some period of time – for trial or pilot evaluation purposes this would ideally be the period of the trial – rather than *denying* treatment, keeping in mind that this delay can also result in a negative customer experience. Many PAs are thus unwilling to conduct a recruit-and-deny/delay study because of concerns about the customer's potentially negative experience. Tension resulting from a customer opting in but then being denied the treatment can also affect the customer's behavior in unknown ways (e.g., attitudinal shifts, interactive effects, other behavioral changes) that may compromise the internal validity of the experiment.

Key RCT Considerations

Randomization not at the individual customer level. Depending on the causal mechanism of an intervention, RCTs can be conducted by randomizing a factor other than the individual, thus avoiding the negative customer experience. For example, if your interest is in getting customers to purchase a certain EE widget and NOT in how they then use that widget, then your focus is specifically on the purchase decision, which can be manipulated at the retail level rather than at the individual level. In this case, randomization could be designed at the store level, rather than at the individual buyer level. For example, half of the stores in a chain could be randomly assigned to receive treatment (for example, through targeted marketing and/or information campaigns, or by offering additional rebates/incentives). The remaining stores assigned to the control condition would simply be "business as usual" approach.

The randomization point (i.e. store/location vs. individual/family) depends not only on the level of causal analysis you wish to investigate, but also on data availability at that level. In the example above, the evaluator would need to ensure that the chain will provide store-level sales data. Other options include random assignment based on geography (zip codes, towns) or time (alternating months of treatment). While these designs overcome some customer experience challenges, they all risk introducing various potential levels of bias that could create noise and possibly compromise the ability to attribute any difference between treatment and control (stores/zip codes/etc.) to the actual treatment.

Randomization checks. While random assignment with sufficiently large samples should yield treatment and control groups that are comparable at the aggregate level, it is advisable to conduct randomization checks to assess the composition and characteristics of treatment and control groups resulting from random assignment. Depending on the design of the study, it may or may not be possible to re-do the random assignment if you find that the groups are not comparable on key attributes. For example, PG&E conducted a recruit-and-deny RCT in which trial participants were recruited face-to-face on a rolling basis over several months. At the point of recruitment, people were screened for eligibility and then informed whether they could participate in the trial or not (based on a computer-generated random assignment.) In this case, during final analysis of the trial, pre-treatment differences in energy use were found between the treatment and control groups, and the final regression analysis needed to be adjusted to account for those pre-treatment differences. In this case, it would not have been possible to repeat the random assignment, but it might have been helpful to know about the pre-treatment differences before the final analysis.

Equity Concerns. An RCT may not be the best approach for a pilot program if the PA has any concerns about not offering this treatment to all of their customers. Some PAs have an equity requirement which prohibits them from offering incentives/rebates/programs – even in pilot form – to anything less than their full territory or customer base. Committing to a full rollout after a pilot, no matter what the results of the pilot may be, may address equity concerns.

Strategic Sampling. It is important to remember that an RCT rests on the assumption of random assignment within the full sample frame, which may or may not be the general population. Because an opt-in program relies on cooperation of those who choose to opt-in to it, they may not be representative of the general population. Thus, in order to maximize external validity, a trial or pilot of that opt-in program, or an evaluation of the program itself, should identify a comparison group comprised of those who are most likely to (or who subsequently do) opt-in for the treatment.

For example, a savings estimation trial for a smart thermostat trial involved sending mailings to a random selection of the PA's customers asking them to respond if they were interested in participating in a contest with a 50/50 chance to win a free smart thermostat that would be professionally installed in their home. The study found energy savings at a different level than in other studies of the same technology, and a subsequent sample analysis showed that the RCT sample was older on average, and in a different life-stage, than the typical product customers. These savings findings are insightful and interesting but, because of the differences between the trial participants and those who are most likely to purchase this product absent an EE program, the resulting savings estimates are unlikely to be comparable to the savings that would result if this PA launched a full-scale rebate program for smart thermostats.

Customer Data for Sampling. When designing the sampling plan for a trial or pilot of an opt-in program, one consideration is whether the PA has reliable customer data to support strategic sampling of the customer segments most likely to participate in a future program. It is ideal to define the sample frame carefully because, with an opt-in program, specific segments may be the most likely to participate in the program. It is critical that recruitment efforts focus on those whose perceptions, actions, and pilot experience mirror the likely adopters once a full-scale program is launched. For example, in a trial for an emerging home management technology, it would be ideal to build a sampling plan using the PA's

customer segments identified as being tech-savvy and gadget-friendly since members of those segments would be most likely to respond to a broad appeal.

Design Integrity. Because the foundation of the RCT design is the premise that customers assigned to the treatment condition are actually treated, and customers assigned to control condition are not treated, there can be a risk in the delay/deny scenarios for an opt-in program if the treatment (or a similar treatment) is publically available. The risk is that a customer who opted-in to the treatment but was denied it went on to find a way to self-treat. For example, if someone opted in to participate in a pilot of a new smart thermostat but was then denied it (or told they had to wait a year), the individual might choose to buy a thermostat independently. It is thus important to monitor your control customers and to build in budget and time to survey them at the end of the trial to find out if (and how) their denial/delay experience affected them and whether the experience affected their behavior, their purchases, or their attitudes. If so, your analysis will need to reflect this. An additional advantage of data collected from the control group is that it could provide valuable insight into changes in the general population or the market for that technology or treatment. This information becomes even more important for long pilots (e.g., 2+ years) and fast-moving markets.

Option 2: Randomized Encouragement Design (RED)

Randomized Encouragement Designs (REDs) are growing in prominence because of their statistical power and compatibility with opt-in programs, but the fundamental requirements of the analysis can limit the situations in which a RED approach is appropriate. Because a RED analysis is highly dependent on disparate uptake levels between the encouraged and non-encouraged groups, the evaluator must realistically consider the program's ability, and budget, to motivate high rates of participation in the encouraged group while limiting participation in the non-encouraged group.

A RED Design

RED is an RCT with encouragement design, with the timing of the randomization being the fundamental difference from a traditional RCT (SEE Action 2012; Cappers et al. 2013). A RED moves the randomization process earlier in the design so that it precedes participation recruitment or customer opt-in. Another key difference is that a RED assumes that not everyone assigned to treatment will end up being treated, and that some members of the comparison (non-participant) group will be treated. This difference requires additional analysis beyond that required for an RCT.

The fundamental premise of a RED design is that everyone in the population of interest is randomly assigned into either an "encouraged" (E) group or an "unencouraged" (U) group. Everyone in the E group is then strongly encouraged (through marketing, communication, incentives, etc.) to participate in the treatment, while everyone in the U group is left unexposed as much as possible so that, ideally, they are unaware of the treatment, do not request it, and do not obtain it independently.

A RED design assumes that the population of interest includes three segments which, because of randomization before recruitment, should be equally represented in both the E and the U groups:

- The first segment is comprised of those who would never opt-in for the treatment, regardless of how much they are encouraged. This segment is referred to in the RED design as the "Never takers" (NT).
- The second segment is comprised of those who will always opt-in, regardless of whether they are encouraged. This group is called the "Always takers" (AT).
- The third segment is the "Compliers" who will opt-in for treatment when encouraged to do so, but will not if not encouraged.

The success of a RED design rests on effective encouragement to convince as many compliers as possible in your treatment group to opt-in for the treatment, and effective non-encouragement so that the only people the U group who opt-in are the Always Takers in that group. To be clear, a RED does not deny treatment to anyone. This is its main advantage over a simpler RCT design. The RED design instead tries to use information and opportunity to "stack the deck" such that many more members in the encouraged group compared to the unencouraged group are aware of the treatment opportunity and are motivated to take advantage of it.

The image below depicts the three RED groups in a population of interest for any particular treatment program: Always Takers are shown in blue, Compliers in Green, and Never Takers in Yellow. After every member of the population of interest is randomly assigned to either the E or the U group, encouragement begins for those in the E group. This encouragement (marketing, recruitment, incentives, etc.) ideally leads the Compliers in that E group to choose to opt in. The Always Takers (blue) in both groups are people who are so motivated that they will have found out about the program and signed up, regardless of any marketing or encouragement to which they were or were not exposed.

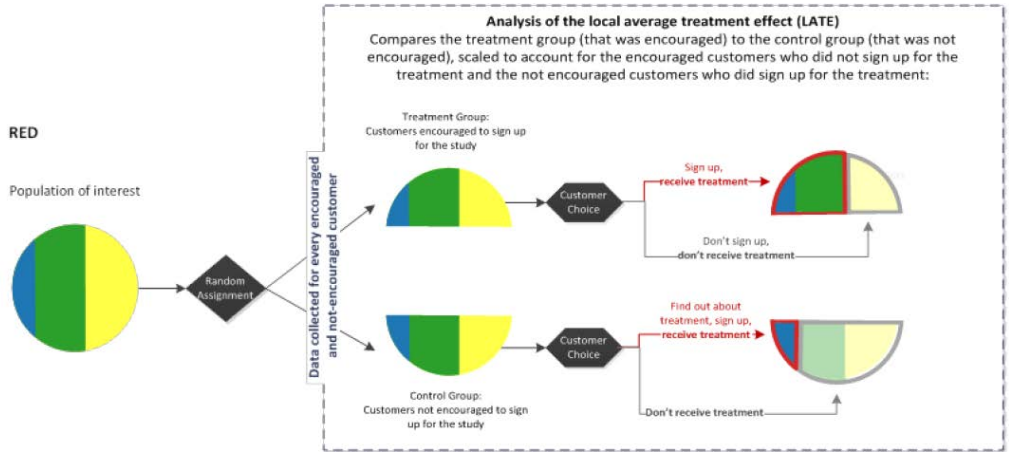

Figure 1. Randomized encouragement design and data analysis. *Source: Cappers et al.*

It is important to note that the type and percentage of AT, NT, and C will vary for each program, and these proportions can only be computed in hindsight, after customers have or have not opted in for treatment. That said researchers should estimate the percentages of each in the population of interest for your program in advance. The percentage of Always Takers in your population of interest can be derived from the opt-in rate in your U group. Assuming the encouragement was effective, the percentage of Compliers in the population of interest can be calculated based on the opt-in rate in your E group, minus the Always Takers. Again, assuming the encouragement was effective, the proportion of Never Takers is the percent of the encouraged group who did not opt-in for treatment.

A RED Analysis

A two-step analysis is required to analyze data collected through a RED. The first step of a RED analysis is the same difference-in-difference analysis one would use to analyze data collected through an RCT. This analysis compares everyone assigned to the E and U groups on the key dependent variable and is called an Intent-to-Treat (ITT) analysis. In a RED, an ITT analysis finds the effect of the "treatment" which, in this case, is encouragement not the effect of the program.

However, because not every member in the E group will have been treated (and some in the U group will have been treated), the ITT analysis of the two full groups is misleading because it fails to

account for the Never Takers in the E group, and the Always Takers in the U group. It is therefore necessary to conduct a second step in the analysis, in order to find the local average treatment effect (LATE) or the effect on treated (EoT). This step involves an adjustment of the ITT effect in order to account for these NTs in the E group, and ATs in the U group. This second step allows us to move beyond the effect of the encouragement so the effect of the treatment on those who were actually exposed to it can be estimated.

RED Considerations

Data availability. Because random assignment in a RED includes the entire population of interest and precedes any customer action or opt-in, the E and U groups tend to be much larger than typical RCT treatment and control groups. For example, it's not unlikely that the population of interest is many thousands, and the E and U groups are each half of that. For the first step of a RED analysis, data from the entire E group, regardless of whether they opted in for treatment, needs to be compared to the entire U group, including those who opted in for treatment despite not being encouraged to do so. The need for data on the full E and U groups can be a challenge in terms of availability, and of the time and cost of procuring and working with so much data.

Opt-in levels. While acknowledging that a RED is never going to be as "clean" as an RCT (in which every member assigned to the treatment group is treated and no member assigned to control is treated), the closer a RED can get to treatment levels of an RCT (presumably 100% treated in the treatment group(s) and 0% treated in the control group), the greater the chance of that RED analysis correctly estimating a treatment effect. The success of a RED evaluation rests on effective encouragement in order to produce opt-in rates as high as possible in the E group, and opt-in rates as low as possible in the U group. Failure to achieve disparate enough levels of treatment uptake between the E and U groups would likely compromise a RED's ability to identify a treatment effect. There are several possible scenarios regarding treatment uptake in the two groups: it could be similar in both (whether it is similarly low or similarly high makes no difference to the analysis but is relevant for freeridership issues and program cost-effectiveness), higher in the U group (though this seems unlikely), or higher in the E group. The latter scenario is necessary to correctly identify an effect, and the higher it is in the E group relative to the U group, the better for the analysis. Fowlie et al. (2015) used a RED to evaluate a program offering free home weatherization retrofits to low-income households. After many thousands of in-person house visits, phone calls, door hangers, and mailed post-cards, the trial achieved a retrofit rate of 6% in the E group compared to less than 1% in the U group. In this case preliminary estimates of savings due to retrofit were 20% of household usage, and it is only because of such a large effect that the analysis was able to identify it with such low uptake in the E group. If a RED approach to an opt-in EE behavior program achieved 6% and 1% uptake rates, it would be almost impossible to reliably identify a typical program effect, given that they are typically much closer to 0% than 20%.

Insight into natural market movement or "the counterfactual". A benefit of a RED is that it provides insight into the counterfactual (what would happen in the market without the program). For example, if there is a new EE product on the market and the program is designed to encourage consumers to buy it, you get a sense of the counterfactual in terms of the purchase rate amongst your unencouraged group. Thus, if 15% of your U group purchases the product, and after encouragement 25% of your E group purchases it, an argument could be made that the market uptake is moving pretty well without an EE program intervention and that the incremental lift provided by a program isn't sufficient to justify its cost. In the event such a program is launched, this counterfactual information suggests that your program could have high rates of free ridership. On the other hand, a 30% treatment uptake rate among the E group, and 2% rate among the U group, strongly suggests that the market is

moving slowly and that an EE program could be a necessary and effective mechanism to accelerate market adoption.

Marketing control. Depending on the nature of the program, it can be a logistical challenge to keep the U group unencouraged. Ideally, those assigned to the U group are unaware of the program and unmotivated to find out about it or participate. However, the need to keep the U group unencouraged (and the E group encouraged) means that any efforts undertaken to inform and motivate the E group need to be skillfully targeted to reach only that group. If the program involves program partners (manufacturers, retailers, etc.) it will be important that they agree not to conduct marketing that could compromise the study design. In addition, any marketing undertaken to encourage the E group will need to be restricted to targeted methods (such as direct mail or email) rather than broad marketing tactics (such as radio, outdoor signage, online ads, etc.), unless broader marketing outreach can be conducted in a way that ensures members of the U group are not exposed (e.g., geographically isolated areas).

Effective Encouragement. A challenge of RED is the requirement for highly effective encouragement, because this can be difficult, time-consuming, and expensive. An key question when considering a RED design is whether there are time, budget, and information resources to do enough targeted marketing to get a high enough uptake rate in the E group. For example, keeping in mind the typically low open/response rates (low single digits) for direct mail, is there sufficient budget to send multiple direct mailings per member of the E group to achieve an uptake rate in the double digits (ideally high double digits)? If one is pursuing an email approach, does the PA have access to enough valid email addresses to reach the numbers required (considering email bounce-back, low open and click-through rates, and the fact that the email recipient then has to be sufficiently persuaded to opt-in)?

Pilot marketing vs. Program marketing. A further marketing challenge for REDs is the fact that a successful RED pilot might require extraordinary marketing in order to achieve the uptake rates needed for a reliable analysis, and that program budgets might not allow for that same type and degree of marketing to be maintained in a future, presumably larger, program. For example, the extensive encouragement efforts described above in the weatherization program (Fowlie et al. 2015) cost more than $1,000 per weatherized household – a cost that most PAs would likely find unsustainable and not cost-effective. An additional consideration is that a RED pilot might require only targeted marketing, which prevents the testing of broad marketing tactics that might be most appropriate for a full program. In addition to marketing tensions, any differences between pilot marketing and future program marketing would mean a pilot uptake rate not representative of the uptake in a broader program rollout – making it hard for PAs to predict program uptake and estimate cost-effectiveness.

Where RCTs and REDs Overlap

While the challenges of conducting a RED might lead a PA to want to avoid that design, the truth is that evaluators often plan to conduct an RCT but end up with a RED because of the real-world difficulties of effectively treating all of the customers assigned to a treatment group. While the shift from RCT to RED is usually unknown problem due to lack of follow-up information on the control group, the fact is that trials of publically-available treatments can be compromised if members of the control group find a way to self-treat (i.e. sign up for the program or buy the technology independently).

When this happens – when an RCT is designed that morphs into a RED due to treatment opt-out or breakage, or because control group participants found a way to self-treat – the RED 2-step analysis can be a reliable way to estimate the treatment effect. Per RCT rules, all customers assigned to the treatment condition would continue to be counted in the treatment group (also termed as the intent-to-treat group), and so the standard difference-in-differences ITT analysis should be conducted and findings reported. However, the RED second step LATE (or EoT) analysis (explained above) should also be conducted in order to more accurately estimate any treatment effect.

Trial Example. A smart thermostat trial was conducted at PG&E as opt-in with recruit-and-deny. Of the 635 individuals assigned to treatment, only 505 actually had the thermostat installed (breakage was due to change of mind, installation no-shows, installation problems, etc.). The RCT ITT analysis that was conducted compared all 635 "treatment" customers to 640 control customers and thus underestimated any effect of the treatment itself because it included 130 customers whose thermostat was never installed. The evaluator conducted a LATE analysis using the 505 homes, in order to specifically assess whether there had been any effect of the treatment on those who actually received the treatment. In this case, none of the ITT (0.7% kWh, -1.4% therms) or LATE (1.0% kWh; -2.0% therms) findings were significant at the .05 level (Churchwell & Sullivan 2014).

Recruit-Under-the-Guise-of-Survey. The Recruit-Under-the-Guise-of-Survey approach is a useful example because it looks like an RCT but is actually a RED and requires the 2-step RED analysis. This approach involves conducting a survey to gather relevant information on a large group (they might be a selection of the general population or targeted customer segments), using that survey data to confirm eligibility for the trial, randomly assigning qualified respondents to treatment or control conditions, and subsequently contacting those assigned to treatment to recruit them to opt-in to the trial. This approach has a number of benefits, including the large data set compiled from all survey responses, avoidance of any treatment denial or delay for the U/control group, and maximizing the chance of a high uptake rate in the E group through pre-screening of respondents based on their survey responses.

Trial Example. PG&E is currently designing a trial for a home energy management technology that is available in stores. Three of PG&E's customer segments are identified as being most likely to purchase this technology independently. A short online survey will be emailed to several hundred thousand customers in these segments. The survey will have questions about their home, home energy management, and whether they might be interested in participating in a trial if one was offered. The population of interest – all respondents who meet the trial's basic requirements (such as being a PG&E dual-fuel customer, no existing energy management technology in the home, etc.) and who indicate they might be interested in participating in a trial – will be randomly assigned into the E or U group. Not all of those assigned to the E group are expected to participate in the trial, but the goal is an uptake rate above 50%. Funds are set aside to conduct follow-up surveys with the E group, but also with a sample of the U group to understand how many of them purchased this technology independently during the course of the trial. The latter are the "Always Takers" who need to be accounted for in the LATE analysis, and whose behavior sheds light on the natural market behavior around this technology.

Option 3: Quasi-experimental Design Approaches

Methods such as propensity score matching, whereby characteristics of the treatment group(s) are used to create matched comparison group(s), have been popular for decades, but are often associated with concerns about the effectiveness of the matching. These concerns increase for evaluations of programs in which there is greater risk of self-selection bias. "*By the very decision to self-select into a program, the members of the treatment group are different from those of any comparison group that can be constructed post-hoc from non-participants.*" (NREL 2013, 8-7) In contrast, some argue that the degree of self-selection bias varies across programs such that matching can be conducted reliably for those opt-in programs where participants and non-participants are not uniquely distinct or where their differences can be identified and measured. Work conducted by Provencher et al. (2013) highlights the fact that matching on long-term energy usage can provide a reliable non-participant control group for evaluating opt-in programs, though this approach requires long-term consumption data that may not be available. Similarly, an evaluation of the California IOUs Residential Audit programs demonstrated highly effective matching that enabled a sufficiently powerful analysis to identify a statistically significant savings estimate of 3% of household usage (Itron 2013).

Glinsmann and Provencher (2013) compared three matching models to evaluate an opt-in behavior program: one model matched participants with non-participants, while the other two were variants of within-subjects participant matching. The resulting point estimates of the opt-in program effect were not statistically different from each other, suggesting minimal self-selection bias, but only the within-participants matching models were significantly different from zero, highlighting the reduced power of non-participant matching and increased risk of type II error.

Variance-in-Adoption. The Variance-in-Adoption (VIA) technique is one of several approaches (for example, The Uniform Methods Project 2013, 4.3-4) to address the challenge of non-participant matching by creating a within-subjects comparison group from the sample of customers who opted-in to the program over time. By matching from within the pool of customers who opted in to a program within a certain period of time (e.g. one calendar year), the VIA approach avoids the challenges of matching to non-participants. Consider two adopters, one opting in to treatment in February 2014 and the second in December 2014. VIA would use the December-adopter's March 2014 data (behavior, usage, etc.) as the control comparison against a February-adopter's treated March 2014 data. While there are constraints around program timing and data availability, the VIA approach can be a good choice if an opt-in program is likely subject to self-selection bias. (See Harding and McNamara's (2011) VIA analysis of ComEd's Energy Saver Program.)

The VIA approach is not applicable for every opt-in program, and there are time weighting factors that must be considered. Two fundamental requirements for conducting a VIA analysis are that the program have a rolling opt-in period, and that there are no inherent differences between people who opted in at different points. A further requirement is that there are sufficient program participants to support reliable matching on other key variables relevant to the behaviors and factors of interest (home size, energy usage levels and patterns, etc.) to afford sufficient statistical power to detect an effect if one exists. Another consideration is that the VIA model has timing implications such that VIA analysis of a year-long opt-in trial could over-weight the data from the earlier treatment months and under-weight data from later months, leading to potential bias. Finally, the design needs to account for market developments that may influence the general population awareness, knowledge, and interest in the program elements to be tested. For example, Apple's purchase of Nest resulted in high-profile marketing campaigns which altered the base conditions of the smart thermostat market. Because of the limitations associated with traditional non-participant matching approaches, and within-subject matching such as the VIA approach, an evaluator limited to a quasi-experimental design should consider conducting both analyses to better understand any self-selection bias (Glinsmann & Provencher 2013; DNV-GL 2014.)

Conclusions

Designing tests to measure the impact of opt-in behavior programs is uniquely challenging due to their typically small effect sizes and their risk of self-selection bias. Reliable evaluation of these programs thus requires powerful evaluation designs without having the luxury of the large sample sizes achieved by the successful opt-out EE behavior programs exemplified in Home Energy Reports.

The critical challenge for using opt-in designs effectively is in creating control groups that are fundamentally similar to the treatment group so that valid comparisons can be made. With careful forethought, the application of the methodologies outlined in this paper can help to address these evaluation challenges: if a PA is amenable to recruit-and-deny/delay, then an RCT is the best approach. If not, the PA should assess whether program design and available budget can support an RED with sufficiently disparate uptake to enable a reliable analysis. If a true experimental approach is not possible, the best alternative may be a quasi-experimental approach with two comparison groups, one of program participants and one of matched non-participants. In all cases, the best design will reflect the program's causal mechanisms, available data, budget, and PA/regulatory tolerance for type I and II errors.

References

Cappers, P., A. Todd, M. Perry, B. Neenan, and R. Boisvert. 2013.*Quantifying the Impacts of Time-based Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies: Protocols and Guidelines.*LBNL-6301E. Berkeley, CA: Lawrence Berkeley National Laboratory and EPRI.

Churchwell, C., and Sullivan, M. 2014. Findings from the Opower/Honeywell Smart Thermostat Field Assessment. PG&E Emerging Technologies Program: ET11PGE3074. San Francisco, CA: Nexant.

DNV-GL. 2014.*Whitepaper: Evaluating Opt-In Behavior Programs: Issues, Challenges and Recommendations* prepared for the California Public Utilities Commission – Energy Division. CPU0088.01 Rev. V1.

Fisher, R. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Fowlie, M., Greenstone, M., and Wolfram, C. 2015. *Are the Non-Monetary Costs of Energy Efficiency Investments Large? Understanding Low Take-up of a Free Energy Efficiency Program*. Berkeley, Calif.: E2e Working Paper 016.

Glinsmann, B. and B. Provencher. 2013. *I can't use a Randomized Controlled Trial – NOW WHAT? Comparison of Methods for Assessing Impacts from Opt-In Behavioral Programs*. International Energy Program Evaluation Conference, Chicago, IL.

Harding, M. and McNamara, P. 2011. *Rewarding Energy Engagement: Evaluation of Electricity Impacts from CUB Energy Saver, a Residential Efficiency Pilot Program in the ComEd Service Territory*. http://smartgridcc.org/wp-content/uploads/2012/01/Stanford_CES-Evaluation_Draft.pdf

Itron. 2013. *2010-2012 CPUC HEES Impact Evaluation Final Report prepared for the California Public Utilities Commission*. CPU0062.01

NREL. 2013. *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. NREL/SR-7A30-53827.

Provencher, B., B. Vittetoe-Glinsmann, A. Dougherty, K. Randazzo, P. Moffitt, and R. Prahl. 2013. *Some Insights on Matching Methods in Estimating Energy Savings for an Opt-In, Behavioral-Based Energy Efficiency Program*. International Energy Program Evaluation Conference, Chicago, IL.

SEE Action. 2012. *Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations*. Customer Information and Behavior Working Group; Evaluation, Measurement, and Verification Working Group. DOE/EE-0734.

Shadish, W., Cook, T., and Campbell, D. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Belmont, Calif.: Wadsworth CENGAGE Learning.