U.S. Department of Energy Smart Grid Investment Grant Technical Advisory Group Guidance Document #7

Topic: Design and Implementation of Program Evaluations that Utilize Randomized Experimental Approaches

November 8, 2010

U.S. Department of Energy Smart Grid Investment Grant

Technical Advisory Group Guidance Document #7*

Topic: Design and Implementation of Program Evaluations that Utilize Randomized Experimental Approaches

November 8, 2010

In any program evaluation context, the quality and usefulness of an estimated intervention effect depends to a significant extent on its internal and external validity. If a study is internally valid, we can credibly state that the estimated impact was caused by the intervention being evaluated. If a study is externally valid, we can confidently extrapolate findings to a larger population of interest.

There is broad consensus in the research community that, under certain conditions, well designed and implemented randomized control trials provide the most valid estimate of an intervention's impact on an outcome of interest.[1] These approaches have been used in a broad range of program evaluation contexts. In this memo, the case for implementing a randomized evaluation of a dynamic pricing intervention is outlined and the practical steps necessary to carry out such a study are explained.[2]

---

* The following individuals on the Lawrence Berkeley National Laboratory Technical Advisory Group (TAG) drafted and/or provided input and comments on one or more of the U.S. Department of Energy Smart Grid Investment Grant (SGIG) Technical Advisory Group Guidance Documents: Peter Cappers, Andrew Satchwell and Charles Goldman (LBNL), Karen Herter (Herter Energy Research Solutions, Inc.), Roger Levy (Levy Associates), Theresa Flaim (Energy Resource Economics, LLC), Rich Scheer (Scheer Ventures, LLC), Lisa Schwartz (Regulatory Assistance Project), Richard Feinberg (Purdue University), Catherine Wolfram, Lucas Davis, Meredith Fowlie, and Severin Borenstein (University of California at Berkeley), Miriam Goldberg, Curt Puckett and Roger Wright (KEMA), Ahmad Faruqui, Sanem Sergici, and Ryan Hledik (Brattle Group), Michael Sullivan, Matt Mercurio, Michael Perry, Josh Bode, and Stephen George (Freeman, Sullivan & Company). In addition to the TAG members listed above, Bernie Neenan and Chris Holmes of the Electric Power Research Institute also provided comments.

[1] See Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." Journal of Economic Literature, 47(1): 5–86.

[2] This guidance document is not intended to imply that the results produced from prior studies of dynamic pricing which did not utilize randomized control trials are invalid. Rather, this guidance document identifies circumstances and situations where randomized control trials would generate results that are at least and/or more precise and credible as those produced under alternative designs.

INCORPORATING RANDOMIZATION INTO PROGRAM EVALUATION DESIGN AND IMPLEMENTATION

In order to obtain internally valid estimates of how an intervention (e.g., dynamic pricing, real time information provision) affects household-level outcomes of interest (such as hourly or daily energy consumption), one needs an unbiased estimate of the household-level behaviors that would have been observed in the absence of the intervention. One approach involves comparing household energy expenditures and related outcomes before and after the intervention. However, this comparison will capture not only the effects of the intervention, but also the effects of other variables that change over time. For example, a before-and-after comparison could under-estimate the effects of a dynamic pricing program if weather were systematically more extreme, or if energy prices were lower and households consumed more, in the year following the intervention.

Multiple regression models can be used to try to control for differences in underlying time trends. There are several reasons why this could be particularly challenging in this context. The best case scenario is that the researcher has access to household-level demographics such as the age and number of household members, employment information, living patterns (e.g., people at home during the day, occupant schedules), as well as detailed information about equipment and appliance ownership (e.g., size, type, number, energy efficiency, and age of different appliances, heating and cooling equipment). However, these household characteristics change over time, and it is unusual to do the kind of in-depth, repeated surveying that would be required to control in the regression for changes over time in these characteristics. Moreover, estimation results may be sensitive to the choice of functional form when modeling the relationship between energy consumption and observed time-varying factors.

Changes in energy consumption at participating households can be compared with changes in energy consumption at a set of observationally similar households before and after the program is introduced. Absent randomization, this kind of non-experimental difference-in-difference (DID) comparison can yield a credible estimate if the treatment effect is large enough and if the comparison of interest is to be made over a short interval of time that unobservable time-varying factors are unlikely to vary substantially (e.g., three months). That said, non-experimental DID designs are not without limitations

Where panel microdata are available, researchers typically improve the fit of their regressions dramatically by including household fixed effects. It is important to emphasize that household fixed effects can control only for time invariant factors. For example, many of the features of the home itself (e.g. type of home, number of floors, outside wall construction material, ceiling height, number of windows, etc.) are largely time invariant and fixed. Concerns arise with DID when there are time varying factors that differ between the treatment and the control group. Households who choose to participate in dynamic pricing programs are likely to have differences in some factors that vary with time along both observable and unobservable dimensions.

Observable differences in time varying factors across treatment and control groups can be difficult to interpret. For example, when households in the treatment group are observed purchasing energy efficient appliances more frequently than households in the control group, is this the causal impact of the treatment or selection (i.e. that these households who chose to participate in the dynamic pricing program are

different)? Perhaps more importantly, the validity of the DID estimates will be undermined if unobserved changes in household energy use over the study period are correlated with the decision to select into the program. A striking finding in electricity regressions is that even after controlling for a rich set of observable characteristics, there are large differences in electricity consumption between households. Even small differences in underlying unobservable trends can confound the effects we are interested in detecting over time. This is more problematic if we are interested in measuring how responses evolve over several months or years. As the time horizon of interest gets longer, it becomes more difficult to know what changes were driven by treatment and what changes were driven by differences in the myriad of other unobservable factors that change over time and impact electricity consumption.

Randomized control trials (RCTs) are widely viewed as the "gold standard" of program evaluation and offer a promising complement to these more standard observational methods. The basic idea is to sample randomly from the population of interest, and then randomly assign selected participants to treatment and control groups. The intervention of interest is administered to the treatment group. The control group, by contrast, receives no intervention and represents what would have happened to the treatment group subjects in the absence of the treatment.[3] The difference-in-differences in observed outcomes across treatment and control groups, before and after the intervention, provides an unbiased estimate of the causal impact of the intervention.

These experimental approaches can be used to leverage both before and after comparisons and comparisons between the experimental treatment and control groups. Direct comparisons of differences in outcomes across treatment and control groups are possible because the effects of selection bias and other confounding factors are eliminated by design. If the study participants have been randomly selected from the population of interest, external validity is also achieved. This means that we can more confidently extrapolate the study findings to the larger population from which the sample was drawn.

Mandatory assignment of households to treatment (or program participation) status across households is not always practical or appropriate for all research questions or contexts. Even if mandatory assignment is possible in principle, it will often be the case that households assigned or offered a treatment (e.g. dynamic pricing tariff) will not comply with or accept their assignment, possibly due to state regulatory policies and practices (e.g., a PUC may decide that customers must make an affirmative choice to opt-in to a dynamic pricing tariff). Given this situation, we highlight a research design that can accommodate these implementation situations: a randomized encouragement design (RED).

THE RANDOMIZED ENCOURAGEMENT DESIGN

The basic idea behind an RED is quite straightforward. The approach involves selecting a subset of eligible households, dividing them into treatment and control groups and then actively encouraging (hence the

[3] In an ideal setting, the control group would be unaware of their participation in the study; however, most practical applications of a dynamic pricing consumer behavior study may require control group participants to be informed as such. Under these circumstances, there are concerns about the "Hawthorne" effect, where individuals in an experiment or study will act differently simply because they are being observed. These concerns should be identified and/or dealt with appropriately in the study design.

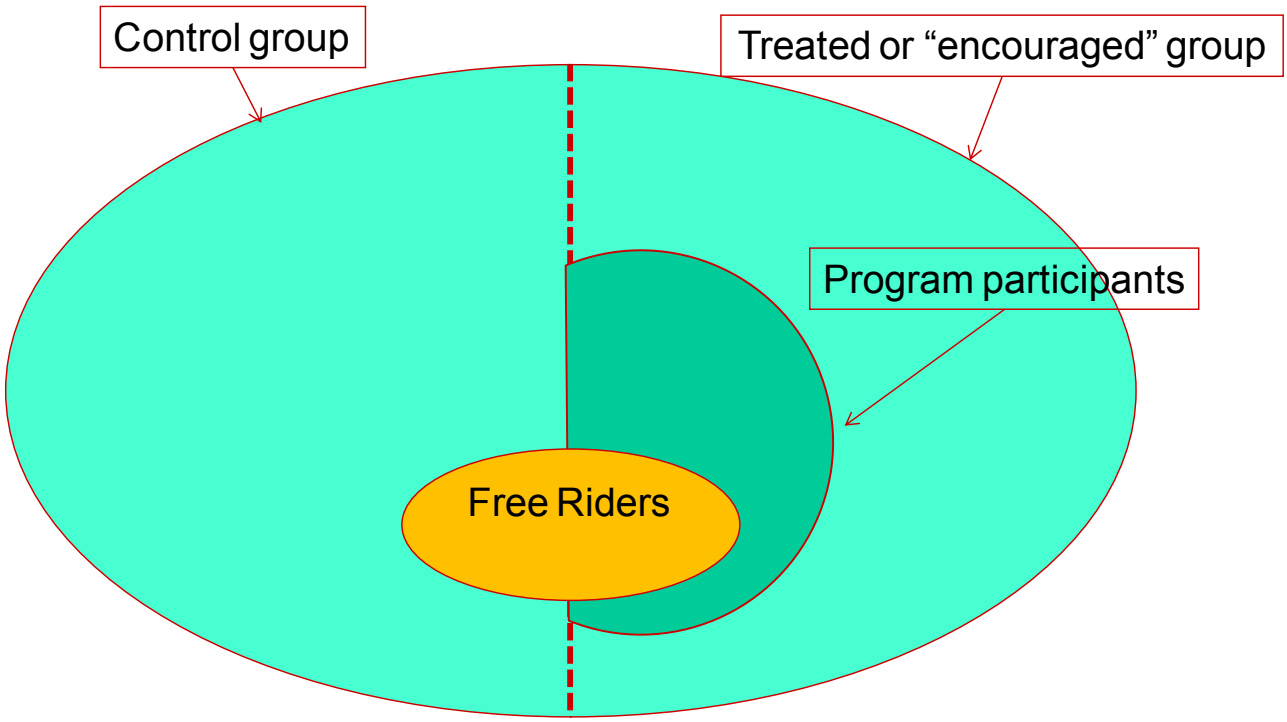
name of the design) households in the treated group to apply for the program. Note that this encouragement can come in many forms. It may manifest, simply, as extending the offer to a household to opt-in to the program or tariff that we are interested in studying. As a result of this encouragement, a larger proportion of the households in the treated group will participate in the program. The analysis proceeds by comparing outcomes across the households who received the encouragement and the households assigned to the control group.

A graphical introduction

Figure 1 diagrams the RED concept. Assume that the large oval represents the sample of eligible house-holds to be studied (e.g., utility customers with advanced metering infrastructure). The first step is to divide the population randomly into a treatment and control group - the two groups will look very similar in every dimension. Importantly, the hard-to-measure characteristics that can be important when interpreting the effects of a program will be distributed similarly across both groups. For instance, both groups will contain similar shares of consumers with strong interests in reducing their utility bills by adjusting their consumption in response to dynamic prices.

The treated group is then encouraged to participate in the program. Some of them will respond to the encouragement, and others will not. In the Figure, the subset of the population that responds is represented with the dark green semi-circle.

Figure 1: Illustration of a randomized encouragement design



With any program evaluation, one crucial issue is to separate true behavioral changes that occur in response to the dynamic pricing program from changes that would have happened anyways. For example, in a program designed to study the effects of critical peak pricing, one could imagine some households signing up for the program because they know they will be on vacation in the hottest summer month, so their consumption will be low in the peak periods (and it would have been low absent the program). In the bubble graph, these households are reflected by the yellow circle. Importantly, there are would-be free-riders (i.e., consumers with the same naturally low peak usage) represented in the control group. So, with an analysis that compares the response of the whole treatment group to the whole control group, customers who would be free-riders if offered the program are expected to be equally represented in both populations. (Note that this also helps explains why it would be misleading to compare the subset of the treatment group who accept an offer to the entire control group drawn randomly from the population. One problem with such a comparison is that (would-be) free-riders would be more heavily represented in the treatment group than in the control group.)

Theoretical foundations of the RED

To illustrate the theory and associated assumptions underlying a randomized encouragement design (RED), we will use notation that is now standard in the econometric and statistical literature. A binary variable D_i indicates whether household *i* has been exposed to the intervention (or participated in the program) of interest (D_i = 1) or not (D_i = 0). Let Y_i denote the outcome observed at household i; for $i = 1…N$. We postulate two potential outcomes: $Y_i(1)$ denotes the outcome that would be realized at household i if it participates in (is exposed to) the dynamic pricing program of interest; $Y_i(0)$ denotes the outcome if household i does not participate/is not exposed. For example, D_i might indicate whether a household i participates in a critical peak pricing program, and Y_i measures household electricity consumption during critical peak events.

Ideally, we would observe both $Y(1)$ and $Y(0)$ for each household. This would allow us to measure causal effect of the intervention for each household (i.e. $Y_i(1)$ - $Y_i(0)$). Household-level measures of impacts could be used to construct not only aggregate impacts of the program, but also estimates of how program impacts vary with observable covariates (e.g., climate, dwelling characteristics, socioeconomic factors). The fundamental problem, of course, is that only one potential outcome can be observed for each household. Thus, to identify the causal effects of the program intervention, an estimate of the so-called "counterfactual" outcome must be constructed. More concretely, if household i participates in a CPP program in time period *t*, we need to estimate what the consumption patterns of household i in time *t* would have been had the household remained in the control state.

In a RED, researchers indirectly manipulate program participation using an encouragement "instrument" so as to generate the exogenous variation in program participation that is so essential for causal inference. This exogenous variation can then be used to identify the effect of the program on those households whose participation was contingent upon the encouragement.

The RED can be explained in the larger context of instrumental variables (IV) methods. Let z_i represent a valid "instrument" for program participation: a variable that is correlated with D_i but uncorrelated with any other determinants of the potential outcomes $Y_i(0)$ and $Y_i(1)$: Let $Y_i(D; z)$ denote the potential outcome at household i if the household has participation status $D_i = D$ and instrument value z_i: Assuming a binary

instrument that takes on a value of either 0 or 1, we denote D_{0i} to be the participation status of household *i* if $z_i = 0$ and D_{1i} to be the participation status of household i if $z_i = 1$.

Identification is predicated on three important assumptions:

A1: Potential outcomes $Y_i(D; z)$ are independent of z_i:

(1) $$[\{Y_i(D,z); \forall\, D, z\}, D_{1i}, D_{0i}] \perp z_i;$$

where $z_i = 1$ if household i is assigned to the actively informed group; $z_i = 0$ otherwise. If assignment to encouraged and control groups is truly random, this assumption should hold by design.

A2: Potential outcomes $Yi(D; z)$ are not directly affected by z_i:

(2) $$Y_i(D,0) = Y_i(D,1) \text{ for } D = 0{,}1$$

If the act of extending the option to participate in a program gets people thinking - and acting- differently, this could introduce bias into the estimates. For example, if households are educated about how stressed the bulk power system is during hot summer days as a means to encourage them to participate in a dynamic pricing pilot, then customers who eschew the offer could conceivably be provided with information that might induce them to turn down their air conditioning during such periods, thereby violating this assumption. Unless there is some expectation that this voluntary behavioral response will be pervasive even though there is very little economic incentive for doing so, such concerns should be substantially discounted.

A3: Monotonicity (i.e. the instrument z_i has a weakly positive effect on program participation for all i):

(3) $$D_{1i} \geq D_{0i}\ \forall\ i$$

Monotonicity implies that the encouragement will never decrease the probability that a household will be exposed to the intervention (although there may be cases where the information or encouragement provided has no effect on program participation). In most, if not all, of the research designs being considered, monotonicity is satisfied by design because control group participants do not have the option to participate in the programs being evaluated. Therefore, the program participation rate in the control group would be zero.

For any given program intervention, consumer behavior study participants can be categorized into one of three non-overlapping groups based on how they react to the encouragement:

1. *Never-takers* are households that will never participate in the program regardless of z_i: Among never-takers, $D_{1i} = D_{0i} = 0$.
2. *Compliers* are households that participate in the program if $z_i = 1$, but otherwise will not participate as they are not formally offered the opportunity ($z_i = 0$). Among the compliers, $D_{1i} = 1; D_{0i} = 0$.

3. *Always takers* are households that will always participate in the program, regardless of z_i.[4] Among always-takers, $D_{1i} = D_{0i} = 1$.

Estimating local average treatment effects

Conditional on assumptions A1:A3, random assignment of information provision allows us to obtain an unbiased estimate of the so-called "local average treatment effect" ($LATE$). The $LATE$ measures the average impact of program participation among compliers:

$$LATE = \frac{E(Y_i|z_i=1)-E(Y_i|z_i=0)}{E(D_i|z_i=1)-E(D_i|z_i=0)} = E\{Y_i(1) - Y_i(0)|D_{1i} > D_{0i}\} \quad (4)$$

Mechanically, our estimate of the local average treatment effect is essentially a weighted average. We construct it by computing the difference in the average energy consumption across the treatment and control groups and dividing this difference by the difference in participation rates across groups. This comparison is meaningful because the proportion of never takers and compliers and always takers will be the same in the treatment and control group in expectation (due to random assignment). Therefore, the contribution of the refusers to the control and treatment group averages, respectively, cancels out in the comparison. All that you are left with is the average treatment effect among the compliers.

This estimand and its statistical properties differ significantly from the average treatment effects estimated using observational methods. First, whereas the conditional independence assumption that rationalizes causal inference in an observational setting is untestable in principle, the independence assumption used to identify (4) is satisfied by design. A second advantage pertains to the construction of confidence intervals. In contrast to observational methods, researchers can remain agnostic about distributions of outcomes and the nature of the underlying sampling process when quantifying uncertainty associated with average treatment effects.

STATISTICAL POWER

In the design stages of any randomized program evaluation, the importance of statistical power calculations cannot be overstated. Whereas randomization can credibly remove bias, these methods do not necessarily remove noise! An underpowered study potentially leads to inconclusive inferences and consequently misspends valuable time and financial resources allocated to the study. An overpowered study may waste valuable resources. Thus, performing sample size and power computations are a critical first step in the design phase.

The power and sample size calculations depend on the planned data analysis strategy. In the context of consumer behavior studies of customer acceptance and/or response to dynamic pricing, the "power" of a study is the ability to correctly detect a difference in group means of a given magnitude. A research design has adequate power if we can be reasonably sure that the observed differences in mean outcomes across

[4] If these “always takers” learn of the program (e.g., from a neighbor who is in the study), they will seek out the opportunity to participate. For simplicity, anyone not offered the treatment should not be allowed to receive the treatment during the study as it may undermine the initial randomization.

treatment and control groups was "caused" by the intervention of interest. More formally, the power of a research design is a measure of the probability of detecting a causal effect of a given magnitude.

Statistical power is influenced by a number of factors and research design choices. This document summarizes the very basics of the power calculations that should be done to inform the design of any randomized field experiment.[5] To be clear, each research design will likely have unique features that will affect how the final power calculation should be conducted. For the purpose of this technical memo, we consider the simple case where we are measuring the effect of a binary intervention (e.g., a critical peak pricing program) on an outcome of interest (household peak period demand). For expositional clarity, we assume we observe each household only once post-treatment. These basic power calculations can be modified to accommodate studies in which household-level outcomes are observed repeatedly pre- and post-intervention. In general, statistical power will increase with the number of observations collected per household.

In the next section, we present simple formulas for calculating power and related statistics. We explain how to do the calculations by hand. Because these calculations are based on the familiar and relatively simple t test, the formulas should be easy to use and understand.

Benchmark power calculation

The basic principles of power calculation can be illustrated using a textbook RCT design in which n subjects are randomly selected from the population of interest.[6] Some proportion p of this randomly selected group of n subjects is assigned to the treatment group and is exposed to the intervention of interest. The remaining $(1-p)n$ subjects in the study are assigned to a control group and are not exposed to the intervention. We assume that all subjects adhere to (or comply perfectly with) their assignment.

In this simple case, the Ordinary Least Squares (OLS) coefficient in a regression of observed outcomes on the treatment indicator provides an unbiased estimate of the LATE:

(5) $$Y_i = \alpha + \beta D_i + \varepsilon_i$$

The variance of the OLS coefficient β is given by

(6) $$var(\hat{\beta}) = \frac{\sigma^2}{\sum(D_i - \bar{D})^2}$$

[5] For more information on power calculations, see Bloom, H. S. (1995). "Minimum detectable effects: A simple way to report the statistical power of experimental designs." Evaluation Review, 19, 547{56}; or Duflo, E., R. Glennerster, and M. Kremer (2006). "Using Randomization in Development Economics Research: A Toolkit". MIT Department of Economics Working Paper No. 06-36 (available at SSRN: http://ssrn.com/abstract=951841).

[6] This textbook example will not cover what would happen in research designs where pre-treatment data is often times readily available. In such cases, the use of this pre-treatment data allows for a reduction in the mean-squared error, which will *ceteris paribus* reduce the necessary sample size, but will also increase the number of total observations thereby reducing the proportion of treated observations, which will *ceteris paribus* increase the necessary sample size. The reduction in the MSE will likely be greater than the effect associated with the reduction in the proportion of treated observations on the required total sample size, but such should be evaluated on a case-by-case basis.

where σ^2 is the error variance. Intuitively, the larger the variation in the unobservables affecting the dependent variable the more difficult it is to estimate β. The denominator measures the variability in the treatment indicator. The more variance there is in this indicator variable, the easier it is to pick up the relationship between the treatment and the outcome. Note that:

(7) $$\sum(D_i - \overline{D})^2 = p(1-p)^2 + (1-p)(-p)^2$$

(8) $$= p(1-p)$$

A simple expression for the variance of our local average treatment effect estimate is:

(9) $$var(\hat{\beta}) = \frac{1}{p(1-p)}\frac{\sigma^2}{n}$$

Ordinarily, one rejects the null hypothesis (i.e., zero effect) when the observed difference between means is large enough such that t exceeds the value set *a priori* to represent the Type I error rate.

Having selected the desired power (minimum probability of detection) κ, size (level at which statistical significance is to be tested) α, proportion p assigned to treatment, and total size of the study group n, the minimum detectable effect can be computed as:

(10) $$MDE = (t_{1-\kappa} + t_{\alpha})\sqrt{var(\hat{\beta})}$$

Rearranging, we can solve for the required sample size given κ, α, MDE, P, and σ^2.

(11) $$N = \frac{(t_{1-\kappa}+t_{\alpha})^2}{p(1-p)}\frac{\sigma^2}{MDE^2}$$

So, in this stylized RCT context, the power calculation depends on the following factors:

- The number of households in the study (N).
- Type I error rate (α). This is the probability of rejecting the null hypothesis when it is true, that is, of incorrectly rejecting the null hypothesis. Appendix D of the Guidebook for ARRA Smart Grid Program Metrics and Benefits recommends an accepted Type 1 error rate α = 0.10.[7]
- The desired level of statistical power (κ). This is the probability that a difference of a given magnitude will be correctly detected. Thus, the power of the test is one minus the probability of not rejecting the null hypothesis when it is false, that is of incorrectly not rejecting the null hypothesis. Power in excess of 0.80 is generally accepted as adequate, but it will depend on the context and the cost of false positives versus false negatives.
- The proportion of the sample receiving the treatment (p). Power is maximized by setting p = 0.5, under a certain set of assumptions.[8]

[7] U.S. Department of Energy, "Guidebook for ARRA Smart Grid Program Metrics and Benefits," Washington, DC, December 7, 2009.

- Minimum detectable (or relevant) effect (MDE). This should be defined as the smallest effect that would justify the program being adopted (versus the expected effect). For example, for a system with a peak demand of 10,000 MW to avoid a new 200 MW peaking generation facility within 2 years, a critical peak pricing program must reduce aggregate peak demand by at least 2%.
- The variance of the outcome. The estimate of variance used in power calculations typically refers to measurement error associated with repeat sampling. This could be the MSE from the regression summarized above.

The power calculations for more complicated research designs are more complicated. The following section discusses one special case in detail.

Power calculation for a RED design

In a RED design, only a fraction of the households in the encouraged group accept (and are exposed to) the intervention. This complicates the power calculation somewhat. Recall that we cannot use all of the variation in the program participation variable D_i to identify the effect of the treatment. We can rest assured that the variation in treatment status of the compliers in the study is independent of the potential outcomes, so this is the variation we will use to identify the treatment effect. Let c denote the share of households that will participate in the program when encouraged.

The OLS coefficient in a regression of outcomes on the treatment (i.e. encouragement) indicator is used to construct the LATE estimate:

(12) $$Y_i = \alpha + \pi Z_i + \epsilon_i$$

Recall that $LATE = \frac{E(Y_i|z_i=1)-E(Y_i|z_i=0)}{E(D_i|z_i=1)-E(D_i|z_i=0)}$, where $E(Y_i|z_i = 1)$ refers to the treated/encouraged group. The point estimate of π is thus divided by the difference in the share of treatment and control group households that participate in the program (c) to obtain an unbiased estimate of the average local average treatment (causal) effect of the program among the households that participate. The variance of this estimator is:

(13) $$var(LATE) = var\left(\frac{\pi}{c}\right)$$

(14) $$= \frac{1}{c^2}\frac{1}{P(1-P)}\frac{\sigma^2}{N}$$

Having selected the desired power, size, P and N, the minimum detectable effect can be computed as:

(15) $$MDE = (t_{1-\kappa} + t_\alpha)\sqrt{var\left(\frac{\hat{\pi}}{c}\right)}$$

[8] In this example, we've postulated a set of model assumptions (Ordinary Least Squares) which means treated and untreated customers have the same variance. There are cases where treatment increases variability, but for now we will not concern ourselves with this possibility.

(16) $$= (t_{1-\kappa} + t_{\alpha})\sqrt{\frac{1}{c^2}\frac{1}{p(1-p)}\frac{\sigma^2}{N}}$$

Rearranging, we can solve for the required sample size given κ, α, MDE, P, c, and σ^2.

(17) $$N = \frac{(t_{1-\kappa}+t_{\alpha})^2}{p(1-p)}\frac{\sigma^2}{MDE^2}\frac{1}{c^2}$$

Note that, as compared to an RCT in which all households comply with their treatment assignment, the number of households required to obtain a given level of statistical power in a RED increases by a factor of $\frac{1}{c^2}$. Thus, for example, if the acceptance rate is 50% among those offered a program, the random encouragement design would require a sample size 4 times as large as the random assignment design, all else being equal. If the acceptance rate is only 10%, a sample size 100 times as large would be needed. If the acceptance rate is only 5%, a sample size of 400 times as large would be needed. This is why the Random Encouragement Design, though unbiased and conceptually straightforward, has practical limitations.