Electricity Markets & Policy
Energy Analysis & Environmental Impacts Division
Lawrence Berkeley National Laboratory

A handbook for designing, implementing, and evaluating successful electric utility pilots

Peter A. Cappers
C. Anna Spurlock

September 2020

This work was supported by the Transmission Permitting and Technical Assistance Division of the U.S. Department of Energy's Office of Electricity under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

Disclaimer

Copyright Notice

A Handbook for Designing, Implementing, and Evaluating Successful Electric Utility Pilots

Prepared for the
Office of Electricity Delivery and Energy Reliability
National Electricity Division
U.S. Department of Energy

Authors
Peter Cappers
C. Anna Spurlock

Ernest Orlando Lawrence Berkeley National Laboratory
1 Cyclotron Road, MS 90R4000
Berkeley CA 94720-8136

September 2020

The work described in this study was funded by the Transmission Planning and Technical Assistance Division of the U.S. Department of Energy's Office of Electricity Delivery and Energy Reliability under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

Acknowledgements

The work described in this study was funded by the Transmission Planning and Technical Assistance Division of the U.S. Department of Energy's Office of Electricity Delivery and Energy Reliability under Lawrence Berkeley National Laboratory Contract No. DE-AC02-05CH11231.

The author would like to thank Asa Hopkins (Synapse Energy Economics), Beia Spiller (Environmental Defense Fund), Bernie Neenan (Energy and Resource Economics), Dave Bisbee and Lupe Jimenez (Sacramento Municipal Utility District), Mary Ann Piette (Lawrence Berkeley National Laboratory), and Tom Stanton (National Regulatory Research Institute) for their review of an earlier draft of this manuscript. We would also like to thank Joy Wang and Anne Armstrong (Michigan Public Service Commission) for the opportunity to vet the ideas included in this report in their MI Power Grid Energy Programs and Technology Pilots workgroup.

Table of Contents

Table of Figures

Acronyms and Abbreviations

EV	Electric Vehicle
IT	Information Technology
LMI	Low-to-Moderate Income
MIPSC	Michigan Public Service Commission
TOU	Time-of-use

Glossary of Terms

Control group – Those not exposed to the elements in the pilot being tested that serve as the counterfactual to those who are exposed to the treatment.
Internal validity – The extent to which one can be confident that a cause-and-effect relationship established in a study cannot be explained by other factors.
External validity – The extent to which one can generalize the findings of a study to other situations, people, settings, and measures.
Null hypothesis – The claim that there is no significant difference in an outcome between specified populations, and that any observed difference is due to sampling or experimental error.
Precision – The extent to which estimates from different samples are close to each other.
Power – The probability of rejecting a null hypothesis when, in fact, it is false.
Self-selection bias – A bias in an outcome that is introduced when individuals select themselves into a group (e.g., participants in a pilot).

1. Introduction

Since at least the late 1970s, electric utilities and their regulators have recognized the value of experimentation to motivate innovation. The industry has a long history of using pilots[1] to help inform future decision making about electric utility rates (Faruqui and Malko, 1983; Caves et al., 1984a; Caves et al., 1984b; Caves et al., 1984c; Hausman and Neufeld, 1984; Lefevre, 1984; Aigner, 1985; Aigner and Hirschberg, 1985; Taylor and Schwartz, 1986), customer technology adoption and integration (Lefevre, 1984; Heffner and Kaufman, 1985; Lalonde, 1986; Yau et al., 1990; EPRI, 1991a, b; Nadel, 1992), and even changes to the utility's regulatory or business model (McCarthy, 2009; Lazar et al., 2011). Utilities have continued to use pilots in this way, especially as they relate to more recent efforts to pursue grid modernization investments, integration of distributed energy resources, and evolution in the utility's business model as part of state-driven or utility-driven innovation initiatives (UtilityDive, 2019; NCCETC, 2020).

Although utility pilots have become almost ubiquitous proving grounds for new rates, technologies, and alterations to the traditional utility regulatory and business model, some regulators are beginning to raise questions about what constitutes a "good" pilot (MIPSC, 2020). There is a sense that some historical pilots have failed to produce actionable outcomes for decision making (Lefevre, 1984; Westlund and Stuart, 2017). Much has been written about utility pilots over the years. Some researchers have sought to understand the level of accuracy or bias produced by the outcomes (Smith and Todd, 2001; Davis et al., 2013; Baylis et al., 2016; Todd et al., 2019). Others have sought to determine what may have caused this bias (Baylis et al., 2016; Todd et al., 2019) or more generally a lack of actionable outcomes (Lefevre, 1984; Westlund and Stuart, 2017). There have been a very limited number of manuscripts providing specific guidance on how to appropriately design and evaluate pilots as part of broader research efforts (Todd et al., 2012; Cappers et al., 2013; Fairbrother et al., 2017; Westlund and Stuart, 2017). [2] However, what is missing from the literature is the identification of a comprehensive process for not only designing and evaluating a pilot, but also implementing, successful

[1] The term *pilot* is often employed in the electric utility industry to mean one of two things. First and most commonly, it refers to an activity undertaken as an **experiment** to determine if something should be pursued more broadly. Alternatively, it refers to an activity undertaken as a test to **ensure success of something** that has already been decided will be undertaken more broadly. An example of the first, as an experiment, would be a pilot where the utility hopes to learn more about customer acceptance, retention, and response to time-based rates from a sample of customers in order to determine if that type of rate design should be rolled out more broadly to the entire customer population. Another example would be a pilot that implements a revenue decoupling mechanism for several years, to see how it functions and determine the financial implications on the utility, shareholders, and ratepayers to determine if is worth implementing more completely and permanently in the future. An example of the second meaning, a test to ensure success of something, would be a pilot that defaults a large share of, but not all, residential customers onto a time-based rate to ensure that all of the utility's back-office systems can handle a future full deployment of all residential customers onto that rate. Throughout this report, we will focus on the former definition, although many of the components of the process discussed do apply to the latter.

[2] Although Sovacool et al. (2018) provides a very comprehensive assessment of how to improve research efforts in the area of energy social science more broadly, they do not include a number of factors that can affect the successful implementation of such research projects, especially as they specifically relate to utility pilots. However, many of the areas they identify which can improve the quality of social science research are highlighted herein.

utility pilots that provide actionable outcomes upon which more informed decisions can be made.

In the sections that follow, a step-by-step process is presented that regulators, policymakers, and utilities can follow to help ensure a pilot is successful, even if whatever is being tested fails to produce the intended or expected result(s).[3] So long as the pilot is implemented as designed and the outcome is determined to meet the necessary level of rigor, accuracy, and precision that subsequent decision making requires, an outcome counter to initial expectations should be viewed as a learning experience, not as a failure. It is best to know that something does not comport with one's *a priori* expectations when implemented on a small scale, rather than implementing something on a much larger scale where the stakes are considerably higher, and then finding out that expectations are not met. Only when the pilot is unsuccessful because of weak design, poor implementation, and/or faulty evaluation should the outcome be considered "bad."

It is worth noting that this report is not intended to serve as a technical resource for those designing, implementing, and evaluating pilots. There are myriad references provided in each section below for those wishing to delve deeper into the technical details. Rather, this report provides a high-level overview of the critical steps one needs to consider when determining the type of pilot one might want to undertake, or the main factors that will have to be thought through to inform the more fine-grained technical details needed for its final design, implementation, and evaluation. Once these high-level critical steps have been accomplished, the technical design and implementation decisions should integrate the perspectives of those individuals and organizations with detailed and comprehensive knowledge and experience associated with the topics covered in each step of the process outlined below.

The report is organized as follows. Section 2 identifies the various critical components for designing a successful pilot. Section 3 focuses on illustrating processes and procedures that should be put in place to ensure successful implementation and execution of the pilot. Finally, Section 4 provides some concluding thoughts and observations.

2. Five Steps of Pilot Design

Five key parameters should be determined before a pilot can be designed (see Figure 1):

1. Identify the *elements* (e.g., outcomes of interest) that will be the focus of the pilot.
2. Decide on the level of *power and precision* needed to satisfactorily identify changes in those pilot elements.[4]
3. Establish the degree of *internal validity* of the pilot's outcomes.
4. Determine the degree of *external validity* of the pilot's outcomes.

[3] Although the report is framed from the perspective of electric utility pilots, the processes and issues are more broadly applicable to any pilot initiative by or for any regulated utility (e.g., gas, water).

[4] There is a glossary of technical terms at the beginning of the report to serve as a reference for those who have more limited experience with statistics and its application in experimentation.

5. *Design the pilot*, taking into account all the decisions made in the prior steps.

Each of these key design essentials is discussed in more detail below. Some may be more relevant than others, depending on the issues to be addressed in the pilot, but there is substantial value in being systematic when designing the pilot.

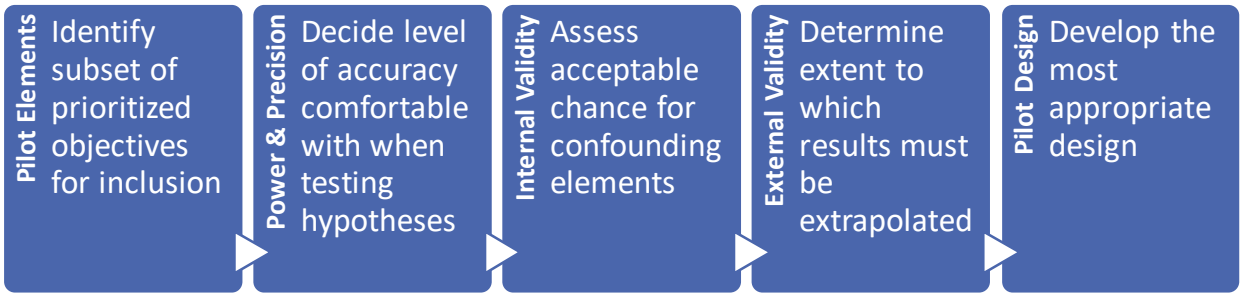

Figure 1. Pilot design process

2.1 Step 1: Identify Key Pilot Elements

The first step in the pilot design process is to determine what issues or topics need to be better understood by the utility, regulators, and possibly stakeholders. This should be substantially informed by a determination among possible alternative future approaches concerning the activity under consideration. A brainstorming session is the perfect venue for developing this set of alternatives, and all ideas should be considered acceptable at this stage of the process. Such brainstorming could take place within the utility or be pursued in a more open forum with stakeholders, regulators, and policymakers providing their input. The time for culling these down to a manageable level comes later; for the goal at this point should be to produce a comprehensive wish list.

For example, suppose there is interest in moving towards broader adoption of time-based retail rates for a residential customer class. A pricing pilot could help all affected parties better understand the myriad of factors contributing to the outcome of interest (see Figure 2). Some of these factors relate to customer acceptance, retention, electricity consumption impacts (e.g., peak demand reduction, energy shifting, flexibility), and bill impacts associated with a transition to a particular time-based rate design. The pilot could focus on determining these impacts for the entire customer population or maybe also for, or only for, particular subpopulations of interest (e.g., low-to-moderate income [LMI] customers, elderly customers, or those with particular medical needs). In addition, the pilot could seek to differentiate these effects based on the enrollment approach onto the rate (e.g., voluntary versus default), or at least take into account expected variation in pilot outcomes based on the alternative enrollment approaches available.

Figure 2. Possible issues for inclusion in a time-based rate pilot

Next a determination should be made about how the learnings associated with each of these issues will be subsequently applied for future decision making. There needs to be a clear linkage between the issue that the pilot seeks to learn more about and how those learnings will help better inform decisions down the road. Absent such a strong linkage, the value of designing the pilot around that issue is reduced, which means the set of elements originally chosen to focus the design on can begin to be culled, to narrow them down to those of the greatest importance and applicability.

For example, suppose high levels of customer retention and satisfaction with a voluntary offering of a time-based rate will be a critical factor in determining whether or not a particular utility decides to move ahead with broad-based implementation of the rate. However, the ability for LMI customers to manage their bills under this time-based rate may determine whether or not the utility will categorically direct these customers away from such a rate offering. There is also some interest in determining whether or not customers will increase consumption during the low-priced off-peak period for planning and operational awareness. But ultimately, this outcome will have little to no effect on the decision to pursue such rate offerings in the future. When taken all together, this combination of observations suggests the pilot should focus substantially on learning about rate designs that are most appealing to all customers but can also be reasonably managed by LMI customers. The pilot should not attempt to learn specifically about issues related to increases in off-peak consumption caused by exposure to this time-based rate.

Once it is clear which questions the pilot should seek to answer, it is worth assessing what others have found concerning that same set of issues. A literature review will help reveal where there have been consistent findings on the issues of interest to the pilot versus situations where outcomes have been more varied or inconclusive (e.g., no outcome has been found at all). The goal of this review is to ultimately determine if the learnings from these other pilots are sufficiently transferable. If they are, then a savvy and cost-conscious utility or regulator could determine that a pilot is not needed to revisit

these particular issues. Effectively, the utility skips the pilot step and uses the results of the literature review to move right into making those decisions which the pilot was intended to help inform. More often than not, however, the differences associated with other pilots (e.g., different retail electricity environment, different location, different customers) overwhelms the viability of extrapolating the results and necessitates the pursuit of a unique pilot.

If it is determined that the utility should move ahead with the pilot, then the list of previously developed critical issues needed for subsequent decision making should be refined to be both more narrowly focused and more explicitly worded into research questions. The more specific these research questions are, the easier it will be to design the pilot to answer those questions.

For example, suppose a critical issue of import is to determine the level of coincident peak demand that can be reduced by exposing customers to a certain type of time-based rate design. A simple research question to address this issue could be:

> *"How much do residential customers reduce their contribution to coincident system peak, on average, during the summertime?"*

However, this does not help produce a focused understanding of what is needed to make subsequent decisions about the pursuit of this time-based rate design as a tool for reducing system coincident peak demand. Conceivably, the reason to pursue this time-based rate is to reduce future installed capacity obligations and purchases in restructured wholesale electricity markets, or to avoid the construction of new peaking generation resources to meet system reliability requirements in vertically integrated markets. Customer response to this time-based rate could be treated like any other resource in wholesale market opportunities or during the utility's integrated resource planning efforts. In either case, there may be some minimum threshold for the level of coincident peak load reduction that should be met to make this a viable option. Conversations with system planners, operators, or the appropriate business process owners could be had and/or analysis could be performed to determine what that threshold is — say, 10 percent. Thus, if coincident peak demand reductions will be an important element in determining the viability of moving ahead with broader adoption of this time-based rate design, the following more refined and focused question will need to be answered:

> *"Will residential customers reduce their contribution to coincident system peak, on average, by at least 10 percent during the summertime?"*

It is likely that the broad list of critical issues will produce numerous highly specific research questions. In fact, a single critical issue of interest could spur the development of many such research questions. Clearly, not all of them may ultimately be included in the pilot, or be the focus for its design. Once a comprehensive list of highly specific research questions has been developed, each should be ranked based on its level of importance and urgency relative to the other research questions (see Figure 3). The resulting set of rankings can be used to determine what to do with each research question with respect to the pilot's design. Those research questions that are of the utmost importance and urgency should

be included in the pilot in some fashion.[5] Alternatively, those that are not as important can be dropped outright if they are not able to be accommodated in the pilot. However, those that are important but need not be answered urgently can be put off and included in some future pilot or research effort.

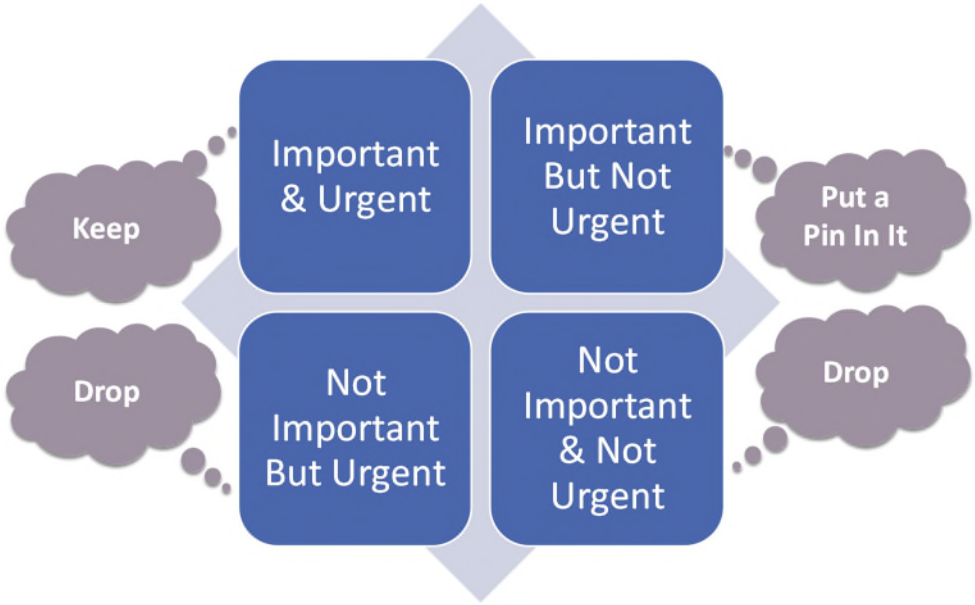

Figure 3. Prioritizing research questions based on importance and urgency

With the subset of critically important and urgent research questions in hand, they should next be translated into testable hypotheses. Technically, a statement worded in such a way that it can either be rejected (because the results suggest or affirm that the statement is false) or not rejected (because the results cannot demonstrate to a sufficient degree of confidence that the statement is false) is called a *null hypothesis.* Statistically, null hypotheses cannot be deemed to be true, so wording them in a way that more readily allows them to be rejected is advantageous from a pilot design standpoint. As with the previous step, this may also result in an expansion of the number of possible null hypotheses that the pilot seeks to test — a single research question could spawn multiple null hypotheses.

For example, one utility designed a pilot to address the following research question (Jimenez et al., 2013):

> *"How does an opt-in time-of-use (TOU) rate without a free enabling technology offer affect participant summer, daily, and event load for residential customers?"*

This single research question was then expanded into three testable null hypotheses as follows:

[5] The concept of urgency should be considered in light of the time it takes to design, implement, and evaluate a pilot that produces results sufficient to make an informed decision; which is likely to be considerably shorter than the time it would take if the pilot needed to provide complete information regarding the same decision.

"During the test period, average daily energy use for residential customers on the opt-in TOU rate without a free technology offer is lower for the treatment group than for the control group."

"During the test period, peak energy use for residential customers on the opt-in TOU rate without a free technology offer is lower for the treatment group than for the control group."

"On event days, peak demand for residential customers on the opt-in TOU rate without a free technology offer is lower for the treatment group than for the control group."

Once all of the critical research questions have been converted into testable null hypotheses, it may be the case that there are simply too many to comprehensively test well within the pilot, given operational or budget constraints. If such is true, then these null hypotheses should be further prioritized, again based on their level of urgency and importance. This activity will ensure that the pilot does not run the risk of being unable to address any of them sufficiently — a situation known as *analysis paralysis.*

It is important to note that some pilots may focus on issues that may not ultimately lend themselves to statistical analysis of these null hypotheses (e.g., implementability or market uptake in utility offerings of rates, programs, products, or services). Testable null hypotheses can and should still be developed. They provide all interested parties, especially those designing the pilot, with a clear focus on what the pilot hopes to learn about, which will subsequently be used to determine whether and/or how to proceed. In addition, they will help inform how the evaluation of the pilot will be undertaken, even if this evaluation entails a relatively simple qualitative assessment of the null hypotheses.

2.2 Step 2: Determine the Required Level of Power and Precision

The next essential step is to determine what level of power and precision is needed when testing these hypotheses. From a statistical standpoint, a pilot with a high degree of power provides greater precision in the estimation of the effect being measured than a pilot with lower power does. Statistical power affects the probability that a test correctly rejects a false null hypotheses. However, the general concepts of power and precision also can be applied in a non-statistical setting to help assess how robust a more qualitative pilot would need to be.

There is a clear trade-off between power/precision and the size/cost of the pilot. Pilots with a high degree of power, and hence precision, are highly unlikely to reject a hypothesis (quantitatively or even qualitatively) when in fact it should not be rejected, while the reverse is true for a pilot with lesser power and precision. A pilot that incorrectly rejects a null hypothesis is ultimately a waste of utility and ratepayer resources, as it results in a pilot that draws erroneous conclusions. However, pilots with a high degree of power will also, all things being equal, require larger sample sizes than pilots with lower

levels of power.[6] A study that collects more data than necessary to answer the identified hypotheses also results in a waste of utility and ratepayer resources.[7] So a balance must be struck when determining the appropriate level of power and precision that a pilot will be designed to have and the effects on the pilot's budget due to the resulting required sample size.[8]

The proper determination for what constitutes an appropriate level of power and precision should consider how the outcome of the pilot will inform subsequent decision making, as well as what the cost would be of coming to the wrong conclusion. If lots of specificity is required or the cost of making the wrong decision is substantial, then the desired level of power and precision should be high, and necessary resources should be provided for the pilot to accommodate the larger required sample sizes. Conversely, if more general information concerning the outcome is sufficient (e.g., direction and order of magnitude rather than more specific point estimates), lower levels of power and precision may be acceptable. In some cases, even determinations that merely address the general concepts qualitatively but do not seek to quantify them statistically may be adequate, resulting in much smaller sample sizes. The outcome of this step will be used, in conjunction with the results of subsequent steps, to derive the exact sample size for the pilot.

For example, if the coincident peak demand reductions induced from the introduction of a time-based rate are going to be used for future resource adequacy purposes, then system planners comparing the cost and performance of these resources against more traditional resources will likely require highly accurate and precise estimates of the average coincident peak demand reductions as well as how consistent these reductions are over time (e.g., the pilot may need to determine, for multiple separate time periods, what the anticipated coincident peak demand would be to within a plus or minus 1 percent margin of error, with a high level of confidence). This means larger sample sizes are needed in the pilot in order to have more power. To achieve greater differentiation when estimating effects for multiple subpopulations (e.g., different demographic groups) or different time periods (e.g., peak versus off-peak, by season) more power and higher sample sizes are needed. However, if the coincident peak demand reduction will not be relied on directly for resource adequacy, then the level of power and precision associated with the estimate can be relaxed (e.g., the pilot's objectives can be satisfied by

[6] Depending on the pilot, a sample could be composed of many different things. For example, a customer who elects to participate in the pilot would be considered a single sample point. In a different pilot, the sample point could be the number of days a product or service offering is available. Drawing firm and accurate conclusions from pilots with smaller numbers of customers or fewer number of days in the field may be more challenging than it would be from pilots with larger numbers of customers or more days in the field.

[7] There are sometimes operational reasons to have sample sizes larger than required for statistical analysis; for example, if the pilot intends to scale into a program immediately if the findings are favorable, or if an operational tool requires adequate testing and is leveraging the research to build use cases for development. It's not uncommon for utilities to leverage the research/pilot period to build the program processes and tools, which may require larger sample sizes to be successful.

[8] In an ideal world these two issues would be jointly resolved. However, utility pilots more frequently start out with a fixed budget, leaving it up to the design team to develop a pilot that fits within the budget constraint. Depending on priorities, this may mean power and precision are sacrificed for a pilot that addresses a more comprehensive set of issues or may result in a much smaller and more focused pilot designed to have greater power and precision to answer a more limited subset of questions.

knowing the result with only a plus or minus 5 percent margin of error). This means sample sizes can be smaller.

2.3 Step 3: Establish the Degree of Internal Validity

Next, it is necessary to consider what constitutes an acceptable chance that confounding effects could distort the outcome of a hypothesis test. Put differently, could something other than what the pilot is trying to test be the cause for the outcome that is observed or derived? In some cases, it is critical that the pilot's design limits, to the degree possible, all opportunities for such confounding effects. These pilots require a high degree of internal validity. Alternatively, it may be that isolating the effect of what is being studied in the pilot is less important, and therefore it is acceptable for outside factors to affect the pilot's outcome.

For example, suppose a utility wants to determine how a particular time-based rate design impacts aggregate electricity consumption. During the same time period of its study there is a promotion encouraging electric vehicle (EV) adoption in the state where the pilot is taking place. If the utility does not track which customers invest in an EV during the pilot (an investment likely to dramatically increase electricity consumption), and the pilot is designed to simply compare the pre-period comparison with the pilot period, then the evaluation of the pilot might erroneously attribute the load growth to the rate instead of to the introduction of an EV. On the other hand, the pilot could be designed and implemented to account for this underlying factor.

2.4 Step 4: Settle on the Degree of External Validity

Aside from internal validity, the design of the pilot can also have consequences for external validity (i.e., extrapolating findings from the pilot to any group, either in the same utility or at a different utility, not included in the pilot itself). As discussed previously, the literature review can help inform if the results found by others could be considered applicable to this particular pilot activity. This is the implication of external validity, enabling the results of a pilot to be extrapolated to a set of customers who did not participate in the pilot. In some cases, it may be very appropriate to extend the results of a pilot to those who did not participate or to circumstances that differ from those in the pilot. In other instances, it may be completely inappropriate to do so.

For example, there may be interest in moving certain customer classes towards a time-based rate that is the default (i.e., customers need to opt-out of the rate in order not to take service under it). Absent any experience with this type of rate under these enrollment conditions, there is interest in measuring customer satisfaction associated with such a transition. However, due to current budgetary and customer service concerns, the pilot has to enroll customers under a voluntary setting (i.e., customers have to opt-in to the pilot to receive the rate). Clearly, the pilot will only measure the satisfaction of those who volunteered for the rate and will exclude anyone who did not volunteer for it. However, in a default enrollment, there is likely to be a subset of customers who would not have volunteered to take service on the rate but do not opt out if the rate is the default (Cappers et al., 2016). The level of satisfaction of this latter group with the default time-based rate is likely to be very different than those

who initially volunteered to take it up, but this cannot be measured in a pilot that only pursues a voluntary enrollment approach. More important, it is these customers that will be the most affected by a transition to a default time-based rates but were never included in the pilot. Thus, if the goal of the pilot is to inform future decision making about default time-based rates, then the results from a voluntary time-based rate pilot are unlikely to have much, if any, external validity relative to that situation.

2.5 Step 5: Determine the Most Appropriate Design

Based on the outcome of the four prior critical pilot design steps, the final design of the pilot can now be completed. However, the design is predicated on one or more research methods that determine how data will be gathered and subsequently analyzed. There are three general "classes" of research methods: experimental, quasi-experimental, and non-experimental. A description of the specific research methods within these broad classes is provided below.

2.5.1 Experimental or quasi-experimental methods

The research methods in this category attempt to test for causal relationships between elements in the pilot, while controlling for external factors that may affect these elements (Sorrell, 2007; Bloom, 2008). Methods that directly employ random assignment of participants to specific treatments of the pilot, as well as to a control group that is not exposed to the elements being tested in the pilot, are classified as experiments. These are considered to be the "gold standard" for establishing causal effects, and therefore have the highest internal validity (Lalonde, 1986). These methods include various forms of randomized control trials or randomized encouragement designs (Campbell and Stanley, 1963; Kirk, 2009). Quasi-experimental methods, in contrast, lack direct random assignment[9], but nonetheless seek to assess the causal effect of pilot elements by controlling for extraneous effects, albeit in a less rigorous manner than true experiments (Price, 2012). Quasi-experimental methods include non-equivalent group designs (i.e., rely on a matched control group) and regression discontinuity designs, to name but a few (Campbell and Stanley, 1963; Reichardt, 2009).

2.5.2 Non-experimental observational methods

In some cases, the pilot cannot or will not be designed to employ random assignment, and a determination is made that it is either impossible or infeasible to control for extraneous effects, yet there is still a strong interest in deriving quantitative results from testable hypotheses (Price, 2012). Such non-experimental methods test for relationships that are correlated but require a set of assumptions, some of which may need to be quite strong, to infer causality (Price, 2012). Because these methods do not control for all relevant external factors, the subsequent analysis of the data generated

[9] This can come about for myriad reasons. For example, a utility may feel uncomfortable denying or even delaying access to elements of the pilot for a subset of participants. So, the utility instead randomizes who qualifies to receive an invitation to the pilot and those who accept the invitation receive the element(s) under study. This approach avoids the challenges of asking customers to participate in a pilot where they may or may not receive the element under study. Alternatively, it may simply be infeasible or impossible to randomize selection or assignment given the elements under study in the pilot.

by the pilot may produce results that include confounding effects, meaning the impacts measured in the pilot may have less internal validity (Dehejia, 2015). Non-experimental methods include correlational designs, descriptive designs, and developmental designs (Radhakrishnan, 2013).

2.5.3 Non-experimental survey methods

If a pilot is focused on assessing the perceptions and attitudes of participants concerning the elements under study, then more qualitative methods are likely to be appropriate. Interviews, focus groups, or survey instruments are all methods for collecting data in a rigorous and structured manner (Radhakrishnan, 2013). Each of these methods seek to gain a deeper understanding of the particular perspective of the respondent, rather than to test an observed action or decision (Sovacool et al., 2018). Although survey methods may be the simplest way to collect data from pilot participants to test hypotheses, they also present some of the greatest challenges when it comes to internal validity and reliability of results vis-à-vis the other research methods discussed above (Coughlan et al., 2009). Poorly designed and/or administered survey methods can generate responses that do not relate to the underlying research questions, are subject to misinterpretation, lack consistency from one respondent to the next, or exhibit significant selection bias, limiting the ability to extrapolate results to the rest of the utility customer base, for example (Umbach, 2005).

2.5.4 Non-experimental case studies

A case study is employed to examine an issue and its associated contextual conditions in depth. Instead of relying on statistical analysis of data from a large sample, case study methods that use quantitative analysis often rely on deductive reasoning to empirically test initial hypotheses (Korzilius, 2012). In contrast, more qualitative case study approaches require the evaluator to understand and interpret what is being observed in the context of the pilot's environment (Korzilius, 2012). Both create opportunities for detailed and comprehensive assessments of innovation, processes, and policies, but create ample opportunity for subjectivity to drive the assessed outcome. In turn, both approaches result in challenges for reliability of the results, as well as their internal and external validity (Hamel, 1993).

2.6 Trade-offs in Key Design Elements and Additional Resources

The decision for how to design the pilot will be predicated on the assessments described above concerning power, precision, and internal and external validity. Experimental methods generally will (or can) have higher degrees of internal validity than quasi-experimental methods, which in turn have higher levels of internal validity than non-experimental methods. Although experimental and quasi-experimental designs can both be used to produce results of roughly comparable power and precision, all three can be implemented to ensure similar degrees of external validity.

For those wanting a more in-depth and detailed discussion of experimental designs and the various issues associated with power, precision, as well as internal and external validity, the following references may be helpful:

- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for*

research: Houghton Mifflin Company.
- Price, P. C. (2012). *Research Methods in Psychology*: Saylor Academy.
- Kirk, R. E. (2009). Experimental Design. In R. E. Millsap & A. Maydeu-Olivares (Eds.), *The SAGE Handbook of Quantitative Methods in Psychology* (pp. 47-72). London, England: SAGE Publishing Ltd.

3. Seven Steps of Critical Pilot Planning

Once the design of the pilot is complete, its success will depend heavily on implementation. Undertaking the necessary level of planning prior to going into the field should substantially increase the likelihood of a successful pilot. What follows is a step-by-step identification of the various types of plans that should be thoroughly completed prior to the commencement of the pilot (see Figure 4).

Figure 4. Pilot planning process

3.1 Step 1: Evaluation Plan

Based on the objectives (see Section 2.1) and design (see Section 2.5) of the pilot, a formal evaluation plan should be developed for the subsequent analysis effort that will be undertaken at one or more points after the pilot begins. However, this evaluation plan should be created concurrent with the design of the pilot. Absence of meaningful feedback and interaction between the design and evaluation efforts risks not thoroughly and comprehensively addressing the critical issues, research questions, and hypotheses that are to be the pilot's focus. It is necessary to determine the appropriate evaluation methods and techniques at the same time that the pilot's design is being undertaken to ensure this problem will not occur. This is especially important because many pilots require data be collected, and

possibly analysis undertaken, prior to the start of the pilot, to develop a baseline for the subsequent evaluation effort.

The evaluation plan should document, at a minimum, three distinct but connected components, as follows:

1. **Establish metrics for testing hypotheses**. The first step in the pilot design process (Section 2.1) concluded by developing hypotheses that the pilot would be designed to test. To test these hypotheses, either qualitatively or quantitatively, metrics must be developed.[10] Thus, it is critical to establish the specific metrics which will be used to test the pilot's hypotheses.
2. **Identify data needs and collection methods**. The metrics which will be used to test the pilot's hypotheses will require data for their construction. The pilot should be designed and implemented so it affords the opportunity to collect the necessary data. This will require not only identifying the data elements needed, but also articulating how, when, and by whom those data will be collected and stored.
3. **Select analytical evaluation techniques**. The collected data will need to be analyzed in order to develop the necessary metrics for hypothesis testing. Several different evaluation techniques probably could be used to analyze those data, but the preferred one(s) should be selected prior to the commencement of the pilot, to ensure the data being collected are consistent with the evaluation technique(s) that will be employed. There is ample opportunity for mismatches to develop between data needs and evaluation techniques, so it is best to resolve these issues before embarking on the pilot to ensure its success on the back end.

All pilots should include some form of an evaluation report that comprehensively documents the design of the pilot, the implementation experience, the analytical methods used, and the results of the analysis, as well how to interpret them. However, the frequency and timing of when results will be reported may differ. Some pilots undertake one or more interim evaluation efforts, where the results can help inform any needed mid-pilot course corrections.

3.2 Step 2: Education Plan[11]

Some pilots intend to test elements that are partially or completely foreign to their participants (e.g., many residential customers have little to no experience with demand charges[12]). In order for the pilots to be successful and have a higher level of external validity, some degree of education may precede the pilot, to ensure the solicitation for participation enables well-informed decision making.[13] Furthermore, there may be value in continuing to educate customers throughout the entirety of the pilot or at

[10] Qualitative (i.e., non-numerical) metrics rely on data that is observational and can be subjective in nature, but can still be used to formally (i.e., statistically) or informally (i.e., inferentially) test hypotheses.
[11] This particular step may not be needed in all pilots. For example, those pilots that do not require informed consent to participate (e.g., piloting a decoupling mechanism) could conceivably skip this step.
[12] A demand charge is a monthly fee applied to a customer maximum's draw of electricity from the grid at some particular time during the month, season, or year.
[13] This education effort may need to be extended to members of a utility's internal staff who will play critical roles in the success of the pilot (e.g., customer service representatives), as well.

strategic points during its execution. To that end, the pilot should have an education plan developed and implemented well before anyone is ever recruited onto it. This education plan should, at a minimum, include the following elements:

1. **Perform an educational needs assessment**. To ensure those who are asked to participate in the pilot are able to make educated and informed decisions, a baseline level of knowledge about the topics or areas that will be piloted should be developed. This baseline can then be compared against the minimum level of knowledge required to make an informed participation decision, to identify what additional level of information and education should be provided.
2. **Develop and implement a pre-recruitment educational campaign**. Based on the educational needs assessment, a campaign to fill the knowledge gap should be developed. This campaign should take into account the likely audience for the educational material in order to determine the most appropriate content and delivery mechanism. This should also consider the necessity of avoiding "confounding" or interfering with the desired pilot outcome to be tested. In some cases, broad-based mass-market educational campaigns may be optimal. In other instances, more micro-targeting may be required to narrowly educate particular subsets of the customer population.
3. **Develop and implement an intra-pilot educational campaign**. It is highly likely that throughout the pilot, participants could benefit from additional educational material to help them be more successful at whatever is being tested in the pilot. This intra-pilot educational campaign could be directly integrated into the design of the pilot, where different delivery mechanisms or content are rigorously tested. Alternatively, simple newsletters or educational material could be provided to inform customers of best practices or basic "tips and tricks" for success, much like that which would be provided to customers during a broad-based rollout of whatever is being tested in the pilot. This should be thought through carefully, however. If that type of educational material is not planned to be a part of a broad-based rollout, then including that material in the pilot may limit the external validity of the pilot, potentially resulting in a larger effect in the pilot than would be expected in a broad-based rollout.
4. **Assess effectiveness of educational campaigns**. Understanding the efficacy of the educational campaigns will not help just with this pilot and what comes after it, but likely will have broader impacts on other consumer engagement efforts by the utility. To that end, the utility should deliberately identify opportunities at different points in the pilot to assess the efficacy of the educational campaign(s). For example, consider adding a soft launch[14] to the campaign that precedes the broader deployment of the educational material by a few weeks. This allows for an opportunity to assess the campaign's effectiveness and make any last-minute minor alterations to ensure its success during the full pilot rollout. Throughout the pilot, it would be wise to get feedback on the efficacy of the educational campaign in order to make modifications as needed. Likewise, assessing how much participants' knowledge has grown

[14] *Soft launches* refer to the release of a product, service, or program to a limited audience to gather information about usage and acceptance in the marketplace before making it generally available to a wider audience at a later point in time.

throughout the education campaign and through the experience of the pilot could be beneficial for future utility efforts.

3.3 Step 3: Marketing Plan[15]

The goal of the education plan is to help customers make more informed participation decisions; the marketing plan is intended to drive customers towards participating in the pilot. The marketing plan should include not just the messages and material that will be used to communicate the opportunity presented by the pilot to potential participants, but also a process for efficiently and effectively onboarding customers into the pilot. Again, the style of these promotional materials should mimic, where appropriate, the marketing materials that would be used in a broad-based rollout of the tested program, to ensure external validity of the pilot. The specific elements of a thorough marketing plan should include the following:

1. **Perform market research to develop a marketing campaign**. Pilots usually test something customers have not had much experience with before but are being asked to take up (e.g., a new retail electric rate or a new piece of technology). The marketing campaign will need to effectively convey why a customer would want to join the pilot and how they could benefit from participating in it by setting appropriate expectations. The best way to develop this marketing campaign is through engaging with would-be pilot participants via market research. Do not assume that you understand your customers and how they might relate to what the pilot is offering. Instead go out and learn about your customers and what messages and channels are better than others at compelling them to join the pilot.
2. **Test your marketing messages and enrollment process**. The market research may suggest certain types of messages and channels will be more effective than others. But to ensure the enrollment process is as successful as it can be, various aspects of those messages and channels should be thoroughly tested before rolling the pilot out. Focus groups and online surveys can be effective tools for testing the marketing material and messages that have been developed. Running a subset of "friends and family" of the pilot (e.g., utility staff) through the enrollment process can help identify ways to improve it. A soft launch of the pilot could be the final test for the marketing materials and enrollment process, as it affords a short window (e.g., two weeks) where small changes can be made before the utility goes out with its full recruitment effort.
3. **Evaluate the effectiveness of the marketing messages and enrollment process.** Although the final enrollment figures will speak to the ultimate success of the marketing messages and enrollment process, it is worth doing a deeper dive into the efficacy (or lack thereof) of these critical elements of the pilot. Such an evaluation requires appropriate processes and procedures for sufficient data collection. The evaluation results will help everyone learn more about what could have been done better, in hindsight, and what learnings can be applied in the future.
4. **Develop and execute an end-of-pilot transition plan**. The goal of many pilots is to learn about

[15] As with the Education Plan, this step in the process could conceivably be avoided for certain types of pilots where customers are not being asked to actively or passively agree to participate.

the long-term viability of whatever is being tested. Depending on the outcome of the pilot, a utility may want to rollout what was tested to its entire applicable customer population or alternatively stop offering it completely. Whatever the next steps may be, pilot participants need to be effectively informed about what happens at the end of the pilot. The last step in the marketing plan should include details about the end-of-pilot transition that allows the utility to clearly communicate to pilot participants, ideally several months before its conclusion, what may be expected of them upon completion and what opportunities exist going forward. For example, in some cases, a pilot will include a piece of technology. Participants need to know what to do with that technology after the pilot is over: Does the utility want it back? Will the utility continue to support, repair, restore, or even replace it? Can it continue to be used in a permanent version of the utility program that was tested during the pilot?

3.4 Step 4: Outreach Plan[16]

In some circumstances, it may be necessary to engage certain types and groups of customers in new and different ways intended to build trust and comfort with the utility, as a precursor to any offerings or actions the utility will undertake. For example, many utilities over the past 15 years undertook substantial outreach efforts prior to their implementation of smart meter deployments, to increase their presence, improve their brand identity and perception, and ultimately build more trust so that their customers would be more open and receptive to the benefits and opportunities presented by advanced meters. Some pilots may also benefit from engaging key stakeholders that include manufacturers, vendors, and contractors. Accordingly, certain types of pilots may necessitate the development of an outreach plan, to improve the likelihood of success. To that end, an outreach plan, if pursued, should include the following elements:

1. **Identify key stakeholders**. The most effective types of outreach target specific groups of customers, whose support will be highly influential in the pilot's success. The first step is thus to identify who these key stakeholders are and to prioritize utility efforts to engage them.
2. **Develop an outreach strategy**. It very well may be the case that different stakeholders will require different approaches to outreach. Once the utility has identified and prioritized these key stakeholders, it should develop strategies for engaging them that take into account their unique characteristics and attributes.
3. **Evaluate the effectiveness of the outreach strategy**. After the outreach strategy has been implemented, if not while the utility is executing it, an evaluation of its effectiveness should be undertaken. The results of this effort will help not just with the pilot for which the outreach was specifically undertaken, but also more broadly for other utility outreach efforts that may be pursued in the future.

3.5 Step 5: Information Technology and Data Management Plan

[16] Depending on the pilot, this step may also be skipped. However, the degree to which any stakeholder engagement is important for the long-term viability of implementing whatever is being piloted may dictate whether this step should be undertaken or not.

Pilots often require a utility to integrate a lot of different information technology (IT) systems that were previously independent of each other or that now need to interact in new and challenging ways. The more IT systems that need to be integrated, the more opportunities there are for problems and failures that may undermine the success of the pilot. This is especially true when the IT systems involved span the utility's back office, out in the field, and even within a customer's premises. Communications networks, computational resources (e.g., bandwidth or storage capacity), firewalls and other security protocols, as well as physical access to the various systems create challenges in the planning, but more important, in the pilot's execution phase. Given the critically important role that IT systems play in the overall success of many pilots, utilities should develop an IT plan that identifies all the systems and their functions that will be leveraged by the pilot, assesses how those systems will need to work together, designs feedback loops where appropriate to ensure problems can be readily identified, and determines solutions for likely problems that may arise.

Data will need to be collected, organized, stored, and analyzed at different points throughout the pilot, as previously discussed (see Section 3.1). Certain types of these data may be considered confidential or proprietary, in some cases necessitating authorization from pilot participants to collect, store, and analyze it, as well as expectations for its protection from unauthorized release. Some data should be accessible in near real-time, while other data will simply be set aside for a future evaluation effort. In either case, having those data available when needed and in an accessible format will be critical for a variety of functions, including troubleshooting when problems occur and evaluating the hypotheses that the pilot intends to test. Utilities would be well served developing a data management plan. This plan identifies all the data elements that should be collected and what they will be used for. It will determine how each data element will be collected, how it will be organized, where it will be stored, and what the data source of record is. Lastly, the plan should illustrate how those data will be accessed and who will have access to it.

3.6 Step 6: Internal Organization Plan

Even some of the simplest pilots touch on many different parts of a utility's organization. They will require access to staff and resources in information technology, customer service, operations, and regulatory affairs, to name but a few. However, pilots, because of their small-scale nature, may be considered lower priority for some of these staff. In fact, their normal, day-to-day tasks are often what they are judged on during annual or semi-annual performance reviews, not their contributions to the success of the utility's pilots, where they may play a minor or tangential role. Because of this, it may be difficult for a pilot manager to gain access to the resources needed to plan and execute the pilot, let alone solve problems in real time that can quickly create both short-term and long-term challenges for the overall success of the pilot.

To ensure that the resources are available when they are needed, it may be beneficial to get the support of senior managers, if not executives, to create cross-functional teams that meet on a periodic basis. This approach not only signals to staff and other leadership the importance of these pilots, but also creates lines of communication and opportunities for a shared vision and mission for the pilot that

should ultimately lead to a greater likelihood of success. For example, these cross-functional teams can help maintain a certain level of situational awareness across the various aspects of the pilot throughout its implementation and execution. If a problem arises in one area, it can be quickly and efficiently communicated to the rest of the team, who can then collectively identify the optimal short-term work-around as well as long-term solution that may require assistance from staff from several different parts of the utility.

Although a formal plan may not always be needed, it is valuable to have some level of documentation that identifies how the pilot will have the necessary access to staff and technical resources at different points during the pilot's design, implementation, and evaluation phases.

3.7 Step 7: External Communication Plan

Many stakeholders, including regulators and policymakers, will be interested in tracking the progress of the pilot. It may be beneficial to develop a plan that lays out how and when the utility will provide updates on the pilot. For example, during the implementation phase, the utility may want to communicate, formally or informally, with a high degree of frequency (e.g., monthly) concerning such things as enrollment statistics, progress towards participation goals, and an assessment of any challenges faced and solutions implemented. Once the pilot is up and running, the frequency with which updates are provided could be reduced (e.g., quarterly, semi-annually) where information about the current status of the pilot (e.g., attrition figures) and critical aspects that will dictate success (e.g., technical problems getting meter data in real-time via telemetry) could be presented. Again, these updates could be provided in written or presentation form as utility filings, but could also be given as part of a less formal discussion or presentation. Finally, the pilot should consider communicating its results, findings, and conclusions in more formal ways (e.g., reports) based on the execution of the evaluation plan. Depending on the duration of the pilot, this should include at a minimum a final evaluation report, but could also incorporate one or more interim evaluation efforts of some kind.

4. Conclusion

Pilots, by their very nature, have a tremendous number of moving parts. Decisions made early on in the design phase may create unexpected challenges during the execution phase. Unanticipated technical challenges may not be identified for months after the pilot gets going, creating billing and IT issues that will need to be resolved expeditiously so they do not jeopardize its success. In the end, all sorts of problems and issues are likely to be faced by the pilot team throughout the full duration of any pilot and its follow-up activities.

Although effective and comprehensive planning can help mitigate the effects of many problems that can be readily identified beforehand, planners will never be able to foresee every challenge nor necessarily identify the best solution to problems ahead of time. Proper planning that includes building coalitions of committed collaborators who frequently communicate should ensure that whatever unexpected challenges do arise will be addressed as quickly and efficiently as possible.

Effective and comprehensive planning should help improve the likelihood of success for the pilot by also increasing the transparency of the various steps in the process. For some pilots, the degree of formal documentation may be much smaller than others, given their limited budget or reduced consequences if the pilot fails to produce results which promote continuation and expansion. In other cases, each of the steps identified may warrant substantial documentation, given the importance of what may be at stake. Ultimately, it is up to regulators, who may need to approve pilots and their budgets for cost recovery, to determine the minimum requirements so that they can make informed decisions that meet their statutory obligations as well as their policy objectives.

The question for regulators and policymakers ultimately is the balance needed between ensuring there is sufficient oversight of the key elements outlined above to make a pilot successful and the limitations that may place on the speed of and utility interest in innovating through the use of pilots. In some cases, a utility may be highly interested and engaged in a pilot, suggesting they will have a greater sense of ownership over its outcomes, potentially resulting in less need for regulatory oversight or shifting the focus of that oversight. In other cases, a regulator or policymaker may push a utility in a direction it is otherwise disinterested in, uncomfortable with, or outright opposed to. Such situations may require substantially more oversight on the part of the regulator or policymaker, and possibly even some sort of positive or negative incentive, to overcome the utility's reticence.

The more regulators and policymakers can create an environment of utility ownership over the pilot's purpose and outcome, the more likely the utility is to support it and do what is necessary to ensure its success. One such approach is a regulatory sandbox (UNSGSA, 2018). This framework can be used by utilities and other electric industry stakeholders to jointly develop, test, and evaluate new or evolving business ideas under limited regulatory supervision without fear of the outcome (OEB, 2020). Regulatory sandboxes encourage innovation through the creation of a relatively low-risk testing environment that promotes experimentation while reducing legal uncertainty regarding the consequences of undesirable outcomes — such is viewed more as an opportunity to learn (e.g., process of invention) than an indictment of poor management (Sheahan and Zhang, 2019).

Regardless of the approach taken, pilots will continue to play a critical role in allowing regulators, policymakers, and utilities to learn from their experiences in order to inform future decision making. Hopefully this manuscript has provided sufficient detail to these entities better understand the key elements that enable more successful pilots.

5. References

Aigner, D. J. (1985) The Residential Electricity Time-of-Use Pricing Experiments: What Have We Learned? Section in Social Experimentation. University of Chicago Press. pp. 11-54.

Aigner, D. J. and Hirschberg, J. G. (1985) Commercial/Industrial Customer Response to Time-of-Use Electricity Prices: Some Experimental Results. *The Rand Journal of Economics.* 16(3): 341-355.

Baylis, P., Cappers, P., Jin, L., Spurlock, A. and Todd, A. (2016). Go for the Silver? Evidence from Field Studies Quantifying the Difference in Evaluation Results between "Gold Standard" Randomized Controlled Trial Methods Versus Quasi-Experimental Methods. Presented at ACEEE Summer Study on Energy Efficiency in Buildings, Asilomar, CA. August 21-26, 2016.

Bloom, H. S. (2008) The Core Analytics of Randomized Experiments for Social Research. *The SAGE handbook of social research methods*: 115-133.

Campbell, D. T. and Stanley, J. C. (1963) Experimental and Quasi-Experimental Designs for Research. Houghton Mifflin Company.

Cappers, P., Spurlock, C. A., Todd, A., Baylis, P., Fowlie, M. and Wolfram, C. (2016) Time-of-Use as a Default Rate for Residential Customers: Issues and Insights. Lawrence Berkeley National Laboratory, Berkeley, CA. June 2016. LBNL-1005704.

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

Caves, D., Christensen, L. and Herriges, J. (1984a) Modeling Alternative Residential Peak-Load Electricity Rate Structures. *Journal of Econometrics.* 24(3): 249-268.

Caves, D. W., Christensen, L. R. and Herriges, J. A. (1984b) Consistency of Residential Customer Response in Time-of-Use Electricity Pricing Experiments. *Journal of Econometrics.* 26(1984): 179-203.

Caves, D. W., Christensen, L. R., Schoech, P. E. and Hendricks, W. (1984c) A Comparison of Different Methodologies in a Case Study of Residential Time-of-Use Electricity Pricing. *Journal of Econometrics.* 26(1984): 17-34.

Coughlan, M., Cronin, P. and Ryan, F. (2009) Survey Research: Process and Limitations. *International Journal of Therapy and Rehabilitation.* 16(1): 9-15.

Davis, A. L., Krishnamurti, T., Fischhoff, B. and Bruine de Bruin, W. (2013) Setting a Standard for Electricity Pilot Studies. *Energy Policy.* 62(2013): 401-409.

Dehejia, R. (2015) Experimental and Non-Experimental Methods in Development Economics: A Porous Dialectic. *Journal of Globalization and Development.* 6(1): 47-69.

EPRI (1991a) Dsm and the T&D System: A Complicated Interaction. July 1991. CU-7394.

EPRI (1991b) Targetting Dsm for T&D Benefits: A Case Study of Pg&E's Delta District. May 1991. TR-100487.

Fairbrother, C., Guccione, L., Henchen, M. and Teixera, A. (2017) Pathways for Innovation: The Role of Pilots and Demonstrations in Reinventing the Utility Business Model. Rocky Mountain Institute.

Faruqui, A. and Malko, J. R. (1983) The Residential Demand for Electricity by Time-of-Use: A Survey of Twelve Experiments with Peak Load Pricing. *Energy.* 8(10): 781-795.

Hamel, J. (1993) Qualitative Research Methods. Section in Case Study Methods. SAGE Publications, Inc. Thousand Oaks, CA.

Hausman, W. J. and Neufeld, J. L. (1984) Time-of-Day Pricing in the U.S. Electric Power Industry at the Turn of the Century. *The Rand Journal of Economics.* 15(1): 116-126.

Heffner, G. C. and Kaufman, D. A. (1985) Distribution Substation Load Impacts of Residential Air Conditioner Load Control. *IEEE Transactions on Power Apparatus and Systems.* PAS-104(7): 1602-1608.

Jimenez, L. R., Potter, J. M. and George, S. S. (2013) Smartpricing Options Interim Evaluation. Sacramento Municipal Utility District. Prepared for U.S. Department of Energy,. October 2013.

Kirk, R. E. (2009) Experimental Design. Section in The Sage Handbook of Quantitative Methods in Psychology. SAGE Publishing Ltd. London, England. pp. 47-72. 9780857020994.

Korzilius, H. (2012) Quantitative Analysis in Case Study. Section in Encyclopedia of Case Study Research. SAGE Publications, Inc. Thousand Oaks, CA. 9781412956703.

Lalonde, R. J. (1986) Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The*

American Economic Review: 604-620.
Lazar, J., Weston, F. and Shirley, W. (2011) Revenue Regulation and Decoupling: A Guide to Theory and Application. Regulatory Assistance Project, Montpelier, VT. June. 94 pages.
Lefevre, S. R. (1984) Using Demonstration Projects to Advance Innovation in Energy. *Public Administration Review.* 44(6): 489-490.
McCarthy, K. E. (2009) Electric Rate Decoupling in Other States. C. O. o. L. Research. January 21. 2009-R-0026.
MIPSC (2020) Utility Pilot Best Practices and Future Pilot Areas. Michigan Public Service Commission - MI Power Grid: Enegy Programs and Technology Pilots Workgroup. September.
Nadel, S. (1992) Utility Demand-Side Management Experience and Potential - a Critical Review. *Annual Review of Energy and the Environment.* 17: 507-535.
NCCETC (2020) 50 States of Grid Modernization: Q4 2019 Quarterly Report & 2019 Policy Review. NC Clean Energy Technology Center. February.
OEB. (2020). Innovation Sandbox. Ontario Energy Board. Retrieved June 15, 2020, from https://www.oeb.ca/_html/sandbox/index.php.
Price, P. C. (2012) Research Methods in Psychology. Saylor Academy.
Radhakrishnan, G. (2013) Non-Experimental Research Designs: Amenable to Nursing Contexts. *Asian Journal of Nursing Education and Research.* 3(1): 25-28.
Reichardt, C. S. (2009) Quasi-Experimental Design. Section in The Sage Handbook of Quantitative Methods in Psychology. SAGE Publishing Ltd. London, England. pp. 47-72. 9780857020994.
Sheahan, B. J. and Zhang, J. (2019) Experiment without Penalty: Can Regulatory 'Sandboxes' Foster Utility Innovation? Smart Cities Dive. Retrieved June 15, 2020. https://www.smartcitiesdive.com/news/experiment-without-penalty-can-regulatory-sandboxes-foster-utility-innov/551012/.
Smith, J. A. and Todd, P. E. (2001) Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods. *The American Economic Review.* 91(2): 112-118.
Sorrell, S. (2007) Improving the Evidence Base for Energy Policy: The Role of Systematic Reviews. *Energy Policy.* 35(3): 1858-1871.
Sovacool, B. K., Axsen, J. and Sorrell, S. (2018) Promoting Novelty, Rigor, and Style in Energy Social Science: Towards Codes of Practice for Appropriate Methods and Research Design. *Energy Research & Social Science.* 45: 12-42.
Taylor, T. and Schwartz, P. (1986) A Residential Demand Charge: Evidence from the Duke Power Time-of-Day Pricing Experiment. *The Energy Journal.* 7(2): 135-151.
Todd, A., Cappers, P., Spurlock, A. and Ling, J. (2019) Spillover as a Cause of Bias in Baseline Evaluation Methods for Demand Response Programs. *Applied Energy.* 250(2019): 344-357.
Todd, A., Stuart, E., Schiller, S. and Goldman, C. (2012) Evaluation, Measurement and Verification (Em&V) or Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations. State and Local Energy Efficiency Action Network, Washington, D.C. May, 2012. 82 pages.
Umbach, P. D. (2005) Getting Back to the Basic of Survey Research. *New Directions for Institutional Research.* 127: 91-100.
UNSGSA (2018) Briefing on Regulatory Sandboxes. UNSGSA's Fintech Sub-Group on Regulatory Sandboxes. Prepared for United Nations Secretary-General's Special Advocate for Inclusive Finance for Development. July.
UtilityDive (2019) State of the Electric Utility Survey 2019. UtilityDive.
Westlund, E. and Stuart, E. A. (2017) The Nonuse, Misuse, and Proper Use of Pilot Studies in Experimental Evaluation Research. *American Journal of Evaluation.* 38(2): 246-261.
Yau, T. S., Smith, W. M., Huff, R. G., Vogt, L. J. and Willis, H. L. (1990) Demand-Side Management Impact on the Transmission and Distribution System. *IEEE Transactions on Power Systems.* 5(2): 506-512.