Supporting and Facilitating Experimentation in Energy Efficiency: Exploring New Validity Constructs to Support Inter- and Intra-Organizational Paths to Success

Anne Dougherty, Mary Sutter, and Katherine Randazzo, Opinion Dynamics Corporation

ABSTRACT

Increasingly, researchers, regulators, and evaluators are looking to experimental design approaches to address two primary concerns in energy programs: insufficient research in the design phase and attribution concerns in evaluation. While there has been much said in the past about how to conduct such studies empirically, few papers have addressed what is required to ensure their successful application for policy and planning.

The authors detail the primary barriers to leveraging experimental design in evaluation through a discussion of validity and the uses and misuses of experimental findings. This is an important discussion, as the energy efficiency industry suffers not from a deficit of talented empirical researchers to design and execute such studies, but rather from a deficit in clear communication and application of this research to policy and planning objectives.

Drawing on select examples in energy and broader evaluation literature, the authors explore the critical considerations needed to ensure that experiments and, more broadly, studies are designed to facilitate planning.

Introduction

Our paper begins with a discussion of the recent call for experimentation in energy efficiency and its history in other disciplines. We then expand on the discussion initiated in the workshop by exploring the role of, and limitations in, experimental design when producing evidence and causal claims for program and policy planning by engaging validity constructs, particularly internal and external validity (defined in the following section). Before we begin, it is necessary to set the parameters for this discussion, specifically: (1) what do we mean when we say "experimentation" and (2) to what end would such experiments be used? To answer the first question, we use the term "experimentation" to refer to experimental design, namely randomized control trials (EPRI 2010). We focus here on the use of experimentation in evaluation, drawing on literature and lessons learned from past evaluations.

The goal of this paper is not to argue for, or against, experiments, but rather to call out that experiments must be better understood and managed in design, implementation and in the application of experiment findings. Throughout this discussion, it is also important to remain mindful that experiments are difficult to implement in field, require careful stewardship, and are suited to treatments that can be highly controlled (such as home energy reports). And it is for these reasons that this paper calls on areas of consideration that should be considered and understood before launching experiments.

Experimentation for Planning: Trading Measurement for Prediction and Planning

Measurement and evaluation in energy efficiency serves multiple purposes. Here we focus on two primary goals of evaluation: (1) retrospectively measuring program impacts, namely energy savings associated with a program, and (2) developing estimates of program savings for resource planning and management.

Increasingly, our industry has been challenged to develop more sophisticated and definitive estimates of retrospectively measuring program impacts. Self-reported attribution is subject to debate, and maturing markets increasingly challenge our ability to tease out the effects of one intervention from the next. These trends, among others, have left the industry wondering, "How can we more effectively measure program impacts?" Experimental designs, in particular randomized control trials (RCTs), have been cited as a potential solution to the industry's challenges. Increasingly, both policymakers and evaluators are calling for more evidence-based evaluation and planning of energy efficiency programs and interventions. In D10-10-033, the California Public Utilities Commission (CPUC) ordered staff to investigate the application of experimental design for energy efficiency programs. In addition, a call for experimentation in energy program evaluation was a frequent topic of conversation at the 2011 International Energy Program Evaluation Conference (IEPEC), and is summarized well by Vine et al. in the following quote:

> *"The use of experimentation, particularly randomized controlled trials (RCTs), has rarely been applied to rigorously test alternative energy efficiency program design features and, more fundamentally, determine the benefits of energy efficiency policy initiatives. The resulting absence of a sound empirical foundation for calculating energy efficiency impacts is impeding progress in the development of effective energy efficiency programs and has led some in the policy community to advocate non-energy efficiency options with more rigorous foundations and less risk of failure in moving the U.S. along a path toward an environmentally more robust energy system (Vine 2011, 2)."*

Recently popularized in the field experiments of behavioral programs, RCTs have demonstrated to regulators, stakeholders, and utilities that they can measure effects of interventions that were otherwise unquantifiable before (such as savings associated with educational reports). As a result, experiments are increasingly the focus of debate in energy program evaluation: Should we (and can we) require greater levels of experimentation in energy program evaluation?

The use of experimentation in other program evaluations has fallen in and out of favor over time. In the late 1960s and early 1970s in evaluation, the federal government mandated experimental designs for the evaluation of education and other social service programs. However, practice fell out of favor due to failure to produce *useful* information (Stufflebeam & Shinkfield 2007; Shadish and Cook 2009). In more recent years, experimental design has focused on producing evaluation of education programs (early 2000s) through the federal mandate for Scientifically Based Research (SBR) specifically focused on RCTs to study educational programs and policies (Christie and Fleisher 2010). However, this recent movement to RCTs is not without its own controversy.

While experiments provide nearly infallible evidence that a program *had* an effect, they are extremely limited in their ability to tell us if an effect *will* occur. As a result, we gain greater precision estimating what *happened in the past*, and often sacrifice our ability to estimate what will *happen in the future*. This is important to call out. As we mentioned earlier, the same impact number in evaluation is used both retrospectively and prospectively. Those who must rely on impact estimates for planning will be ill served by experiments that are not designed specifically to help *predict* and *plan*.

Validity constructs provide individuals and organizations a framework to determine whether a given study effectively measured what it claimed to measure. The two forms of validity that are most commonly looked to are internal and external validity:

- *Internal Validity* is defined by the question "did...the experimental stimulus make some significant difference in this *specific instance*" (emphasis ours).

- *External Validity* is defined by the question "to what populations, settings, and variables can this effect be generalized?" (Campbell 1957, 297 as cited by Shadish, Campbell, and Cook 2002, 37).

Internal validity assures us that the effects did occur (what happened) and external validity advises us as to how these effects should be interpreted and applied (or not) to other situations, times, or populations (what might/will happen). We discuss these throughout this paper.

Internal Validity, the Seductress

Internal validity is highly valued because it assures the researcher that impact findings are "real," e.g., we feel confident they actually occurred. True experiments offer relatively infallible evidence (as compared to other methods) that a treatment effect *occurred.* Due to the high levels of certainty that experiments produce, impact findings gained through experiments are often taken with a level of certainty reserved for laws of physics – if *x* caused *y, x* will always cause *y*. In this way, internal validity is seductive; it provides a level of certainty of a given intervention's impact that is rare in energy efficiency evaluation. Because the causal inference *for the specific conditions of the experiment* are difficult to refute, the limitations of the experiment's findings *under other conditions* are often overlooked in our search for more definitive savings values.

Yet to pull off a successful experiment, researchers have to impose high levels of control over the test condition; treatments have to be clear, manageable, and controlled and treatment populations have to be well understood and relatively static. As a result, it is challenging to apply the findings from experiments to real-world circumstances where programs (treatments) are subject to change and the people and organizations treated are in a constant state of flux. As Gargani and Donaldson state:

> *"The traditional framework is intended to help researchers generalizing the results of experimental research, not stakeholders predicting the performance of programs in specific, local contexts. The former can be a proxy for the latter, but an imperfect one because treatment protocols…are tidier, narrower, and more replicable than programs." (Gargani and Donaldson 2011, 24).*

To restate the point of the authors, in most cases, experimental conditions are controlled and specifically defined. Each experiment is designed drawing on a specific and carefully contemplated sample frame, conducted at a particular moment in time, and with a well-defined (and most often unchanging) treatment. In fact, it is this specificity and relative control that produce internal validity and enhanced certainty in causal inferences. Such tidiness and control over interventions are rare in energy program models, even more so as markets continually mature and transform.

External Validity, the Diamond in the Rough

External validity, in contrast, assures us that our study can be applied to other populations. It offers researchers, policymakers, and program planners a framework for understanding how the research findings might be *applied*. If we acknowledge as an industry that we must use impact findings for planning, then external validity should be given equal, if not more, weight when determining which methods to use in our research.

In their work "What Works for Whom, Where, Why, for What, and When? Using Evaluation Evidence to Take action in Local Contexts," Gargani and Donaldson discuss how both policymakers and research seek to understand "what works" in general terms and as a definitive conclusion, without recognizing (or clearly articulating) the extent to which "what works" is/was determined by the conditions of the experiment, or in their terms, the "local context," that is rarely the same from one program initiative to the next. They state:

> *"'what works' does not explicitly ask for a description of the past performance of a program (what worked?) or a prediction of future performance (what will work?). Rather, it requires that the past and future be addressed simultaneously [and] suggests a false line of reasoning – to know past performance is to know future performance." (2011, 19)*

In energy efficiency, we are susceptible to this line of reasoning. In many cases, such reasoning is a matter of course; evaluations are designed to identify what worked for a given treatment, at a specific moment in time, and for a specific population (ex post analysis of savings), and these findings are used to determine what will work (e.g., as deemed or ex ante savings estimates).

Internal validity is not external validity; the likelihood that the energy savings identified in a given experiment will be reproduced is enhanced when the conditions of the experiment, specifically the persons/subjects, settings, treatments, outcomes, and time(s) (Cook 1993) are similar, central considerations for external validity. For this reason, studies must be set up with external validity in mind if they are to be used for planning and prediction, or complemented by other research, such as formative research, market characterizations/assessments, etc.

One of These Things is Not like the Other: What Happens when External and Internal Validity are Confused

As noted earlier, behavioral programs have introduced the largest RCT experiments ever conducted in energy efficiency. As part of the program model, energy report programs randomly

assign customers to treatment and control groups to facilitate a rigorous evaluation of program savings. When considered as a whole, these Home Energy Report programs are the largest field experiments ever conducted in energy efficiency.

According to the Environmental Defense Fund (EDF), on a national basis, these Home Energy Reports (or information-based energy efficiency programs) have driven individual household savings ranging from an average of 1–3% per year (EDF 2011). These findings were derived from data collected from 11 different utilities and include more than 750,000 U.S. households. Our own research has demonstrated these initial savings on the order of 1.6–1.8% savings (Dougherty 2011).

The study's regression results indicate that the savings potential of this program may be greatly enhanced when targeting based on baseline usage and other variables. When doing so, the savings gained through the Home Energy Report increases, ranging from 0.9% savings when no targeting is applied to 12.8% savings when targeting is applied using multiple variables (Davis 2011, 9). This indicates that the target population may have a significant impact on the savings generated through the treatment, underscoring the point that the specific conditions and subjects of an experiment interact with treatment effects.

Utilities that were early adopters of behavioral programs relied on program savings estimates drawn from initial pilot studies to estimate their savings goals. In a few cases, this generated poor planning assumptions, as the original test region of the pilot was quite different from the target region (and people) of the program. In some cases, savings reflected the higher end of the program's potential (in the 3% range, say, for example, in Colorado) and others the lower end of the savings potential (0.8–1% range, say, for example, in Arkansas).[1] When the latter was true, this meant that some programs fell well below goal.

In the cases where results fell short, a number of factors contributed to the inappropriate application of these ex post savings assessments to planning, such as a lack of additional studies or a reliable body of research to demonstrate this potential variation in savings. In addition, planners may not have sought the advice of, or were not adequately cautioned by, evaluators on the ability to generalize these findings. If, for example, the planners had been engaged in an exercise to determine the study's external validity, or if the evaluators clearly articulated the limits of the evaluation in the findings, better planning estimates may have been derived from the original experiments.

It's Alive! How Studies are (Mis)Used in Practice

As researchers, it is easy to forget that our work "lives" beyond the first and last pages of our reports. As a framework, validity is a communication tool that is used to convey the merits of a study. In this way, validity enters into the policy and planning world as a sort of currency through which the value of a study is communicated. Gargani and Donaldson aptly detail how validity is used in this way. They describe validity as both an argument and a warrant to illustrate the ways in which validity is engaged in practice. Validity as an argument is the most obvious use of validity; researchers use validity to make the case that their research findings can be used in a particular way for a particular purpose. However, we (practitioners and policymakers alike) use validity as a warrant; we use the evidence derived from a study for a given purpose by citing a validity argument (conspicuously or not) to justify the way we apply our findings. When

[1] Note that we obscure the locale of these savings estimates to avoid calling out any one program.

considered in this light, validity is not a neutral construct to objectively assess the accuracy of a study, but rather a living currency exchanged between practitioner and policymaker and, with this engagement, a shared responsibility between the practitioner and the study user-to-user (Gargani and Donaldson 2011, 20).

In pursuit of science, rigor, and the respect of our peers, we researchers create tomes that detail our methods and findings, with endless appendices of data collection instruments and tables for others to recreate our work. What we fail to include, however, is a discussion of the potential uses and applications of our study, which may be the greatest disservice to our research and our clients. In most cases, the uses and potential misuses are outlined in presentations, discussions, stakeholder meetings, etc. and are rarely documented as standard practices. Yet to be responsible stewards of our research, and to avoid the challenges described in the previous section, threats to validity should be made clear.

Reading the Tea Leaves: Determining When to Use Studies for Planning

How can we avoid inappropriately using impact findings for planning? When we consider external validity, there are a number of questions that researchers, policymakers, and planners can ask themselves when leveraging impact findings for planning. Specifically, those interested in using the study have to determine whether the impacts, and the causality behind the impacts, will apply in other settings, e.g., whether the findings are generalizable. Shadish, Campbell, and Cook provide a grounded theory for making causal generalizations based on experimental research findings. They offer five instructive principles of generalization that can serve as guidelines for study users when seeking to generalize findings:

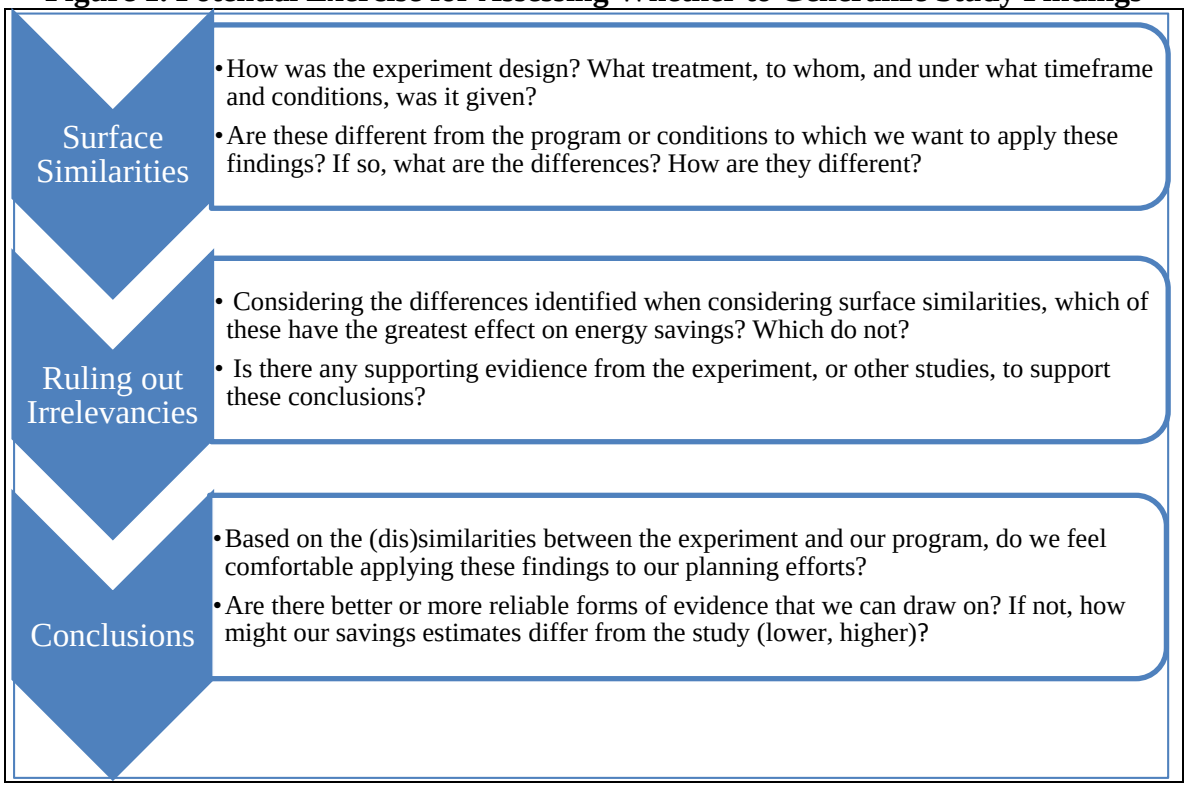
1. *Surface Similarity.* Assessing the apparent similarities between [the] study and the characteristics of [the] population to which we seek to generalize.
2. *Ruling out Irrelevancies.* Identifying those things that are irrelevant because they do not affect or change our ability to generalize.
3. *Interpolation and Extrapolation.* Making interpolations to unsampled values with the range of sampled instances and explore extrapolations beyond the sampled ranges.
4. *Marking Discriminations.* Clarifying the primary discriminations (characteristics) that limit or confine generalization.
5. *Causal Explanation.* Developing and testing theories to discern patterns in effects, causes, and meditational processes that allow for the transfer of causal relationships (closely paraphrased [Shadish, Campbell, and Cook 2002, 24–25]).

If used as a sort of checklist, these five criteria can serve as the basis for determining when the application of study findings are warranted or should be scrutinized. We propose that this framework be used in two capacities: (1) as a teaching tool for policymakers and stakeholders (as well as those new to evaluative research), and (2) as a standard assessment of experiment findings when reporting on their results.

When used as a teaching tool, surface similarity and ruling out irrelevancies may be the simplest concepts to teach and are relatively intuitive assessments of a study. If framed as questions, arriving at the answers may be a relatively simple exercise or something that can be done as standard practice in planning meetings. For example, if study users want to assess surface similarities, simple open-ended question and laddering techniques would facilitate a

critical engagement with the research and, ideally, the arrival at an instructional answer. Figure 1 below provides a high-level overview of how such an exercise might be conducted.

Figure 1. Potential Exercise for Assessing Whether to Generalize Study Findings



For the second goal, we propose that researchers, in presenting experiment findings, engage all five criteria for generalization as standard practice for communicating the potential utility of the study findings to its readers. At best, this will serve to educate the reader and provide the necessary cautions to prevent the misuse of the study findings. At the very least, it serves a means to "cover-off" on the evidence, speaking to Gargani and Donaldson's call for the responsible application of validity as argument and warrant (Gargani and Donaldson 2011).

Begin as You Mean to Go On: Setting Up Research for All Intended Uses

Considering the two earlier stated goals of evaluation: (1) retrospectively measuring program impacts, namely energy savings associated with a program, and (2) developing estimates of program savings for resource planning and management, we must be clear from the onset which of these goals is most important. Both of these goals aim to measure energy savings, which on the surface seems to be a common one. Because this goal seems a common one, it is also frequently reasoned that the same energy savings value can be applied to meet both the objective of knowing something about the past and predicting something about the future. But as we discussed, measuring what happened is particular to the conditions of the study and the program (e.g., its time, place, location, and subjects) and cannot be confidently used to determine what will happen if the exact study conditions are not met in the future. For this reason,

researchers and policymakers must be clear on the intended use of the experiment to ensure that the study is designed to meet its stated goals.

Editors Chen, Donaldson, and Mark recently published a volume in the American Evaluation Association's New Directions for Evaluation Series, *Advancing Validity in Outcome Evaluation: Theory and Practice (2011)*. The editors identify a core challenge in valuing internal validity above all other needs. In their introduction to the volume, the authors write:

> *"According to the Program Evaluation Standards (Joint Committee on Standards for Education Evaluation, 1994), four attributes are essential for the evaluation practice: utility, feasibility, propriety, and accuracy. The Campbellian typology offers clear strengths in addressing accuracy. However it is less suited to address issues of utility, propriety, and feasibility."*

For most studies, accuracy is not the be all and end all. Studies must also be designed to be used, remain ethical, and be feasible to produce and reproduce. For policymakers and planners, the utility of a study must be clearly defined. In most cases, a study is not successful if it does not adequately address the goals of its funders and stakeholders by providing credible results in support of these goals. Yet defining the goals of a given study is more challenging than one might think. Often, researchers conduct studies to answer a set of research questions that, while defined through input with study stakeholders, fail to meet the goals of stakeholders. This occurs primarily when goals are broadly or inadequately specified, and studies are conducted in good faith to answer one set of questions when stakeholders really sought to answer a different set. Conflating verification and measurement needs with planning needs is chief among the sins of vaguely stated goals. For this reason, it is important to begin studies with a clear understanding of how they will be used, essentially beginning as you mean to go on.

Take, for example, the use of pilot studies to develop program models. In their truest form, pilots are tests with the same goals as experiments: to test the outcomes of a given treatment on the intended subjects. Pilots should, at their best, explore the effects of a treatment on all possible applications of the treatment under expected/standard program circumstances. All subjects (participants) of all varieties (homeowners, low income, etc.) and under all conditions (seasons, regions, etc.) that may be present in the implementation program should be included or replicated in the pilot study. In this way, the findings of the pilot can be more confidently generalized to the program conditions (external validity). This can be done smartly by ensuring that the sample frame is representative of the future treatment group, and that the treatment and treatment conditions (seasons, requirements) align with the program design.

However, pilots are often hastily designed or designed with the wrong goal in mind. For example, in efforts to pass cost-effectiveness tests, pilots are designed for customers that are most likely to save; however, a full program rollout most often includes customers who are not likely to save. As a result, what was tested was whether a given treatment *was* cost effective, not whether the treatment *will be* cost effective. Or, pilots are used to see if a treatment will net energy savings under the best circumstances, forgetting that the best circumstances are not often present in full program rollouts.

Further, such limitations on the pilot design thwart the experimental nature of a pilot; when designed to meet objectives other than experimentation and discovery, we design our pilots to measure what we think is true, rather than using them to determine what is true. In both cases, we fail to design experiments that can be used for planning, and limit our pilots to testing

questions that are relevant to our immediate needs and goals, in lieu of discovering strategies that may enhance the long-term success and viability of a program.

To overcome these challenges, researchers and study funders have to align on the study objectives, with a careful eye to how the study will be used. This can be done through deliberate and focused questions and planning at the onset of a research initiative.

Conclusions

Experimentation has its place in energy program evaluation. As with other publicly funded program, the evaluation of energy programs should be held to the same standards of demonstrating program effectiveness and causality. In this respect, experiments have their place in energy program evaluation and have been demonstrated to serve as an effective tool for enhanced impact analysis.

As the industry continues to discuss the potential application of experiments in energy program evaluation, we must remain clear of the *utility* of each specific study and refrain from treating experiments as infallible panaceas that meet all research needs. Rather, we must acknowledge that experiments must be designed with an eye towards their intended uses, and the findings derived from experiments should carry the same caveats and cautions other evaluation methods with similar (and perhaps more obvious) shortcomings.

Further, we also have to consider that accuracy in estimating ex post impacts is one of many goals for program evaluation, and increased accuracy in planning and forecasting is another potentially equally important goal. As we have conveyed throughout this paper, to look back is not the same as to look forward, and we must design our studies with each objective in mind. For this reason, external validity must made an equal partner when we tell the story of our findings, and "live with" our research as it is used and reused for policy and planning purposes.

In all of these ways, we must acknowledge how research is used; for this reason evaluators, program planners and policy-makers alike should consider ourselves careful stewards of our findings and their application and caution against haste and hubris when conveying or applying our results. This is true for all research, experiments notwithstanding.

References

Chen, HT., S.I. Donaldson, and M.M. Mark, editors. 2011. *Advancing Validity in Outcome Evaluation: Theory and Practice*. *New Directions for Evaluation* (130) (summer).

Christie, Christina A. and Dreolin Nesbitt Fleisher. 2010. "Insight into Evaluation Practice: A Content Analysis of Designs and Methods Used in Evaluation Studies Published in North American Evaluation-Focused Journals." *American Journal of Evaluation* 31 (3) (September).

Davis, Matt. 2011. "Behavior and Energy Savings; Evidence from a Series of Experimental Interventions." Environmental Defense Fund. http://www.edf.org/news/study-concludes-in formation-based-energy-efficiency-can-save-americans-billions.

Donaldson, S.I., C.A.Christie, and M.M. Mark, editors. 2009. *What Counts as Credible Evidence in Applied Research and Evaluation Practice?*: Sage Publications.

Dougherty, A.E. 2011. *Massachusetts Cross-Cutting Behavioral Program Impact Evaluation, Volumes I and II.*

Dougherty, A., M Sutter., and K. Randazzo. 2011. "Experimentation in Energy Efficiency: Lessons from the Trenches." Paper presented at the California Public Utilities Commission, San Francisco, Calif., October 24.

[EPRI] Electric Power Research Institute. 2010. "Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols." Palo Alto, Calif.: Electric Power Research Institute.

Gargani, J. and S.I. Donaldson. 2011. "What Works for Whom, Where, Why, for What, and When? Using Evaluation Evidence to Take Action in Local Contexts." Advancing Validity in Outcome Evaluation: Theory and Practice. *New Directions for Evaluation.* (130) (summer).

Shadish, Campbell T. and Thomas D. Cook. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton Mifflin Company.

Shadish, William R. and Thomas D. Cook. 2009. "The Renaissance of Field Experimentation in Evaluating Interventions." *Annual Review of Psychology* 60: 607–29.

Stufflebeam, Daniel L. and Anthony J. Shinkfield. 2007. *Evaluation Theory, Models, & Applications*: Jossey-Bass.

Sullivan, M. 2009. *Using Experiments to Foster Innovation and Improve the Effectiveness of Energy Efficiency Programs*. Berkeley, Calif.: California Institute for Energy and Environment.

Vine, E., M. Sullivan, L. Lutzenhiser, and C. Blumstein. 2011. "Experimentation and the Evaluation of Energy Efficiency Programs: Will the Twain Meet?" International Energy Program Evaluation Conference (IEPEC), Boston, Mass.