One Good Trial:
Experimental Evidence in Electricity Behavior Research

Alexander L. Davis*, Tamar Krishnamurti, Baruch Fischhoff, Jay Apt

Department of Engineering and Public Policy
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213

Abstract

We propose a practical and achievable "gold standard" for electricity field trials. We discuss its content and rationale, providing justifications for why each element is important, then detail how field research looking at electricity consumption can be designed so that the standard is met. We present a detailed example of a hypothetical in-home display field trial to demonstrate feasibility, along with a discussion of the costs of adhering to the standard. We conclude with arguments for why federally funded research on human behavior and electricity consumption should be required to meet this standard, just as the FDA requires this high standard of evidence for drug approval.

Keywords: Field Trials, Research Methods, Reporting, Electricity Behavior

1. Introduction

Providing residential electricity customers information about their electricity use on a custom in-home display can yield small but important reductions in aggregate residential electricity demand.[1] Other interventions that aim to reduce residential electricity consumption may have similar effects. These interventions benefit customers in the form of lower electricity bills, and benefit utilities by reducing "incremental capacity, transmission, and distribution investments".[2] Perhaps most importantly, reducing demand curtails rapidly increasing carbon emissions, striving to reach the goal of limiting global temperature rises to 2° Celsius in the next century.

Without reliable evidence of effectiveness, money will be wasted on interventions that either fail upon implementation or limp on without any

*Corresponding Author. Phone: 1-412-216-2040. Email: alexander [dot] l [l] davis1 [at] gmail.com

real understanding of their usefulness.[3] Well-designed experiments produce reliable evidence by unambiguously establishing both the magnitude of an intervention effect and whether it is causal. Poorly designed ones often yield conflicting results that require repeated (and costly) data collection to sort out the facts.

Because poorly designed experiments will likely suffer from common flaws and, as a result, common bias, the aggregate knowledge produced is not much better than what is learned from a single study. Policy makers should be reluctant to accept such evidence, as a single flawed study can easily support the incorrect conclusion that an ineffective treatment is effective, or vice versa. Even when the safety or efficacy of a new treatment appears obvious to the utility, such as the implementation of smart meters, a single scientifically rigorous study could quell potential customer backlash. One good trial, carefully designed to minimize bias, can provide evidence that is invaluable because of its accuracy, resulting in diminished long-term costs to the utility from continued "pilot" studies, and the implementation of only those technologies and programs that most benefit customers.

The FDA routinely makes decisions about whether to accept evidence of the effectiveness of a new drug or medical device in the form of phase I–III clinical trials. These trials are structured similarly to trials on electricity consumption behavior, so the FDA's standard of evidence serves as a useful guideline for the electricity industry to adopt. At first glance the focus and outcomes of FDA trials, such as cancer treatment and lives lost, may seem incommensurable with the outcomes of electricity field trials. However, it is precisely because the outcomes of medical clinical trials are so imperative that the necessary time and thought has gone into creating a pristine, but also easily translatable, approach. FDA's evaluations ultimately hinge on having two "gold standard" double blind, placebo controlled, randomized trials showing effectiveness. We show how this same approach can be applied to electricity field trials to gather the right evidence for an ultimately lower long-term cost.

In this paper we argue that the FDA's standard of evidence should be adopted by all electricity industry decision makers, from public utility commissions to the Department of Energy, when deciding whether to invest in new technologies or implement behavioral interventions. We discuss the standard of evidence the FDA accepts, giving examples of problems and solutions from biomedical research, and then make the case for applying this standard to research on electricity behavior, using an in-home display field trial as an example. Finally, we provide a simple checklist for designing research that meets the gold standard.

2. The Gold Standard

In this section we discuss the important elements of clinical trials that are relevant to interventions that affect electricity consumption, drawing on government standards (e.g., The International Conference on Harmonization,[1] and the NCI,[2]) texts on the topic[4,5], and quality standards (e.g., GRADE[6]).

Specifically, we propose that a type of experimental approach used in clinical trials, called a *pragmatic* or *effectiveness* randomized controlled trial (RCT), matches the needs of electricity research in terms of both cost and practical application.[7] Pragmatic trials use large representative samples with minimal exclusions,[3] often comparing established treatments against each other to find the best one, such as massage against acupuncture for treatment of lower back pain.[8] Pragmatic trials are used in biomedical research in circumstances that resemble those faced in electricity research, where diagnoses are clear (e.g., a need to cut peak load), conditions are common (e.g., load patterns are predictable and apply to most customers), practical advice on best interventions is needed (e.g., which pricing program or technology is most effective), and interventions are easily implemented (i.e., they don't require technical experts to implement).

We acknowledge that it may not always be easy to implement pragmatic RCTs. Using heterogeneous groups and a representative sample means that the trial requires a large sample size and as such, may have a higher cost than a less rigorous trial. The large sample size also makes quality control and the use of very precise measurements difficult, as precise procedures and quality assurance do not scale well. However, even in the face of these limitations, there is rarely a case when non-random non-controlled alternatives (currently the most common approach to electricity intervention testing) are a better option than an RCT for gathering adequate evidence.

Pragmatic trials can provide the highest quality evidence on the effectiveness of a new technology or intervention, but the evidence that is produced is only as good as the soundness of the study design. In the next few sections we discuss the elements of the study design that are necessary to conduct an ideal successful RCT.

[1] http://www.ich.org/products/guidelines/efficacy/article/efficacy-guidelines.html

[2] http://www.cancer.gov/clinicaltrials

3. Design

3.1. Background, Rationale, and Systematic Review

The design of an RCT begins by establishing the background and rationale of the study. This is facilitated by formulating study objectives into testable hypotheses, specifying measurable benefits and risks.

A systematic review then looks at published and unpublished scientific research on the topic. Theoretically plausible and empirically supported mechanisms of action may be discovered from the review, allowing the study design to be refined as to be able to detect these mechanisms. The review also includes a quantitative meta-analysis that allows calculation of the expected effect size, power, and required sample size of the trial, taking into account the methodological quality of data.[9]

3.2. Internal Validity

A study's internal validity is the degree to which it can be used to make causal inferences. This approach formalizes the intuitive notion of causality, which holds that an intervention is causal if it was not otherwise affected by uncontrolled variables, only affects the outcome (and not vice versa), and works as intended.[10] In the past twenty years causality has been precisely defined in the mathematical formalism of graph theory, making statements about causality clear and well-defined.[11,12] To establish internal validity, RCTs use a concurrent control group, randomization, and blinding.

3.2.1. Control Group

To demonstrate causality, one must simultaneously show that the same participant benefits when given the treatment and does not benefit when not given the treatment. Unfortunately this is impossible, as the same participant cannot simultaneously receive and not receive the treatment. One commonly used approach in the electricity industry is to apply statistical methods to predict a customer's future consumption from past behavior. This approach is unlikely to work, as no statistical model can account for unforeseeable changes in circumstances or unique historical events that may make the future quite unlike the past.

An RCT approach uses a concurrent control group that gives some participants the treatment and a separate group of participants, sampled in the same way from the same population during the same time period, no treatment. If each participant is equally likely to be in the control or treatment group, in the long run the control group should accurately predict what those in the treatment group would have done had they not received the treatment.

The causal argument from an RCT is delicate, as causality is established only if the control and treatment group are treated exactly alike. If, on the other hand, those who receive the treatment know they received it, whereas those in the control group know they didn't, then their beliefs will differ, introducing non-equivalence. This is why placebo groups are used in RCTs, as they ensure that participants cannot tell whether they received the treatment or control, thereby equating the beliefs of those in each group. It is for this reason that special care must be taken to make the placebo as identical to the treatment as possible. For example, a pilot trial of dynamic pricing would need to notify participants in the placebo group of peak time hours, even while charging them a fixed rate per kWh. Creating the appropriate study recruitment language to temper expectations in the placebo group is challenging, but certainly not insurmountable.

3.2.2. Randomization

The only way to make sure each participant is equally likely to be in the control or treatment group is to use a randomizing device, such as a pseudo-random number generator. This prevents the participant and experimenter from determining what group the participant will be assigned to by adding unpredictability into group assignment,[13,14] in the long run balancing variables other than the treatment between the groups.[15] With the best intentions a researcher may, consciously or unconsciously, be tempted to place a larger household with the potential for greater benefit in an in-home display treatment group and a single-person smaller household in the control. Similarly a participant may only select to sign-up for a trial during the period in which the most personally appealing offer is available.

Two critical aspects of implementing randomization are *Sequence Generation* and *Allocation Concealment*. Sequence generation is the process by which randomization is done, for example using a pseudo-random number generator.[16] Random sequences are necessary for group assignment because they are unpredictable, preventing the participant and experimenter from determining the participant's group assignment. Non-random methods, such as alternating assignment or assignment by birth date, allow the possibility that participants enroll in such a way that they receive the treatment of their choice. Using a simple randomization method, such as a pseudo-random number generator, is an easy way to avoid accidentally using non-random sequences that are perceived to be random.[17]

Allocation concealment prevents the researcher and participant from knowing the random sequence and thus whether the next participant will be assigned to the treatment or control group. The allocation must be concealed to both the researcher and participant, as researchers have been known to

decipher and subvert random assignment out of curiosity or the desire to "help" the participant.[18–20]

Here is a concrete example of how an allocation concealment plan may work. Suppose an in-home display trial is conducted where the researcher administering the trial has two possible boxes to mail to a participant, one with the letter A on it and one with the letter B on it. This researcher calls a central facility (for example, the in-home display vendor) and asks whether to give the participant box A or box B, not knowing which contains the fully-functional in-home display and which contains the placebo device (details on what might comprise a placebo in-home display device can be found in section 4.2.3). The central office then generates a random number assigning box A if the number is odd and box B if the number is even. The assigned box (A or B) is then recorded by the central office, given to the participant, and is then collected at the end of the study to make sure it was correctly administered. If the vendor is in charge of pushing information to the in-home display, the researcher may never need to be unblinded to the participants' condition. This procedure prevents the researcher from subverting the treatment assignment and does not allow the participant to decide whether to continue participation in the study with knowledge of which treatment she received. Failing to adequately implement and report allocation concealment can severely undermine the study's internal validity, as there is evidence that studies that lack allocation concealment are consistently biased in favor of effectiveness.[21]

3.2.3. Blinding

Merely knowing that one is in the treatment rather than control group can change both the way a participant behaves and the way they are treated, making the need to hide the treatment assignment to participants and researchers paramount. If the hypothesis being tested is telegraphed, participants may consciously or unconsciously behave in a way consistent with that hypothesis.[22] For example, if a participant is aware that they are in a study with a goal of consumption reduction, they may use less electricity than usual, not because of the actual intervention, but because they want to do what they think they are supposed to do in the study. A striking example of this comes from one of the first blinded experiments. Benjamin Franklin, Antoine Lavoisier and others gave "unmagnetized water" to those who believed "animal magnetism" applied to water could cure illness. Franklin and Lavoisier observed the same hysteric, and supposedly curative, response to both magnetized and unmagnetized water.[23] Without using a blinded experiment, the French government may have concluded that animal magnetism was an effective treatment.

To qualify as a "double-blind" trial, participants and those who interact with the participants or the data must not know the treatment assignment of any participant.[24] However, the definition, use, and implementation of "blinding" is so poorly agreed upon that some have proposed removing it from experimental terminology.[25] For example, medical studies report little or no information about what is meant by "double blind" in reports of their trial,[24] and less than half of studies that could use double blind report doing so.[26]

Maintaining blinding is difficult because unpredictable events and haphazard experimental design can unblind participants or researchers.[27] Contacts (e.g., peak time notification), visits (for equipment installation), and treatment adjustments may convey information about group assignment and thus unintentionally unblind or otherwise make groups non-equivalent. Therefore, all actions done with the treatment group are also appropriately mimicked with the placebo group.

3.3. External Validity

Internal validity is the degree to which causal inferences can be made about the intervention in the study sample. However, RCTs aim to address causal effects in the population of interest, not just the sample, which requires an assessment of the study's external validity, or the degree to which the sample is representative of the population. If the sample was taken randomly from the population, then the causal effect in the sample is an unbiased estimate of the causal effect in the population. While the practical constraints on trials will inevitable produce an imperfect sample, minimizing problems with *Exclusion Criteria*, *Volunteer Adjustment*, and *Withdrawal Prevention*, should allow for a sample that is as close to truly random as possible.

3.3.1. Eligibility and Exclusion Criteria

Pragmatic trials define the population of interest as all people who may potentially benefit from the intervention, attempting to impose minimal exclusion restrictions on this population. Because including such a heterogeneous population in the study can increase variability in the data, the sample size needs to be larger to obtain adequate statistical power. The advantage is that the sample obtained from this approach is representative of the population and thus externally valid,[3] meaning recommendations for any member of the population can be made because they could have been in the sample.[28]

Some exclusions are almost inevitably necessary, however. For example, women would neither provide information on the effectiveness of a prostate cancer treatment, nor could they possibly benefit, and it would be unethical to expose them to an intervention that had any risk of adverse events without

the potential to benefit them or provide scientific knowledge. Similarly, in an electricity intervention, it would not make sense to give an in-home display to the homeless. Eligibility and exclusion criteria avoid this situation by specifying who can participate in the study and who cannot.

These exclusions should ideally be strongly justified, but in practice are often based on weak justifications. A strong justification for exclusion may be the participant was unable to consent, may be harmed by the treatment, or that the intervention may be confounded for that participant (e.g., by cointerventions such as being on both a time-of-use and flat-rate tariff). Weak justifications are based on unmotivated socio-demographic or health factors, such as age, sex, IQ, or chronic condition.[29]

3.3.2. Volunteer Adjustment

Although pragmatic trials use minimal exclusions, the decision to participate in the trial is ultimately up to the participant. In most trials those who participate are volunteers who may systematically differ from those who choose not to volunteer.[3,30] For example, those with higher education, higher socioeconomic status, and women are more likely to volunteer to participate in research studies than those who are less educated, have lower SES, or are men.[31]

Recruiting participants into trials is difficult, with some able to recruit only 2-3% of those offered to participate.[32] The success of the recruitment depends on a number of factors, such as whether the recruiter is comfortable discussing the uncertainty in risks and benefits of trial participation, whether the participant perceives their role in the study as being a "guinea-pig",[33] trust in the researcher, and time constraints.[34]

Best practices for recruitment based on systematic reviews are available[35] and consistently show the effectiveness of using an "opt-out" design, where participants are assumed to want to participate in the study unless they explicitly refuse study participation, rather than fail to respond. This is supported by evidence showing that those who fail to respond to recruitment do so not because they are refusing to participate, but because of other reasons such as not getting around to it or being misinformed about the study.[36,37]

The RCT takes measures to maximize recruitment and accommodate volunteering when people refuse to participate. One approach to accommodation is to use propensity score approaches that model each participant's probability or "propensity" to volunteer.[38,39] Such a model may be developed from psychodemographic variables collected on the recruited population. If the model is accurate, then statistically controlling for propensity to volunteer allows one to make valid inferences from sample to population.

3.3.3. Withdrawal Prevention

Retaining participants in a study is just as difficult as recruiting them.[40] Some of those who agree to participate in the study may not remain in the study until it is complete. If those who are not benefiting from the study also withdraw, then average treatment effects observed in the study will be biased toward showing a greater benefit than there really is.

One approach to accommodate withdrawals is to measure primary outcome variables shortly after baseline to capture critical information before withdrawals occur.[41] For example, a critical peak pricing program should plan to call several critical peak days soon after the study begins. Another approach uses appropriate exclusion criteria during recruitment that accurately predict whether people will be able to complete the study.

Categories or "themes" of retention strategies include community involvement, study identity, training personnel, study description, contact and scheduling methods, reminders, visit characteristics, study benefits, financial incentives, reimbursement, non-financial incentives, and tracking methods.[42] Using multiple strategies across multiple retention themes may be effective.

3.4. Statistical Validity

RCTs stick to basic statistical analyses, for example looking at treatment effects on the whole population, before turning to more complex models and comparisons, such as preplanned subgroup analyses. Once participants are in the study and randomly assigned to condition they are analyzed according to this condition assignment regardless of whether they adhered to the treatment regimen or dropped out of the study.[43] Alternative approaches that attempt to "guess" at missing data, called imputation,[44] can complement the original intent-to-treat analysis.

Careful attention must be paid to the timing and choice of statistical analyses, as experts on clinical trials report that failing to appropriately conduct and report statistics is the most common form of trial misconduct.[45] A representative example is the case where one statistical test produces a p-value greater than .05, a different test (with different assumptions) produces one lower than .05, while covariate adjustments increase or decrease the p-value in arbitrary ways. When a situation like this arises, unblinded analysts are likely to report only the approaches that produced statistically significant ($p < .05$) results.[46–48]

One way to maintain the validity of statistical analyses is to register the trial[3] and publish the protocol of the study ahead of time, including the plan

[3] http://clinicaltrials.gov/, http://www.who.int/ictrp/en/

for statistical analyses, making clear what statistical tests are planned and what are post-hoc. This protocol includes the statistical methods that will be used, the endpoints that will be compared, the target sample size,[49] a statistical power analysis,[50] and clear rules for early stopping.[51] Blind data analysis can complement the protocol publishing approach, but researchers must be wary of pressure that may be put on blinded analysts if results do not come out as predicted.[52]

3.5. Reporting

The Consolidated Standards of Reporting Trials (CONSORT) is a reporting standard that includes all critical information necessary to evaluate the validity of a clinical trial.[53,544] It covers elements that are usually reported, such as the hypotheses of the study, as well as those that are frequently overlooked, such as changes to the protocol after beginning the trial, eligibility and exclusion criteria, sample size determination, the method of sequence generation, how blinding was done, how similar the treatment and placebo were, the allocation schedule and code-breaking, and an evaluation of the success of the blinding.

Along with meeting the requirements of the CONSORT statement, careful attention is paid to selective reporting. Selective reporting occurs when a measure or hypothesis that is not statistically significant is not included in a published report.[55] Selective reporting is common and can undermine the validity of the study report, making treatments look effective when they are not, as all of the negative data are omitted from publication.[56–59]

To avoid selective reporting, RCTs commit to data sharing and reproducibility. Data sharing means providing data to others so they can review the data, check for errors, and reuse it for secondary analyses. This can be done in a variety of ways, including using online databases such as Harvard's Dataverse.[5] Ideally, enough information will be provided about the methods, design, and statistical analysis so that the trial can be replicated and the statistical analyses verified by independent third parties.[60] In those instances in which data may be commercially sensitive–a realistic challenge with most utilities–measures can be taken to anonymize individual data and even the specifics of treatment conditions, so as to merely allow for independent replication of analyses. For example, participants can be assigned a numeric identifier and treatment and control conditions can merely be labeled as "A" and "B," allowing subsequent researchers to determine if there

[4]The CONSORT checklist: http://www.consort-statement.org/index.aspx?o=2965

[5]http://thedata.org/

are outcome differences between the two groups without knowing any greater detail about them. To facilitate reproducibility of the statistical analyses the code can be shared, for example using a combination of the open-source softwares R, LaTeX and Sweave.[61] To facilitate reproducibility of the trial itself, the complexities of the study can be reported in open lab notebooks, such as Open-Wetware.[6]

4. An In-Home Display Trial Example

In this section we provide an example of a hypothetical but highly realistic in-home display trial. This trial incorporates elements of trials that have been conducted in the field, making it both practical and achievable. However, it has been modified to meet the standards the FDA would accept. It is reported according to the CONSORT statement.[7]

4.1. Introduction and Systematic Review

Researchers have extensively studied three approaches to reducing demand-side electricity consumption: dynamic pricing, in-home displays, and automated control systems, such as smart thermostats. A meta-analysis aggregating thirty-two studies on the topic, weighting each study by their precision and adjusting for plausible methodological bias,[1] suggested that in-home displays were the most promising approach to reduce overall consumption, whereas dynamic pricing and smart thermostats were effective for reducing peak but not overall demand. None of the studies on in-home displays were randomized double-blind placebo controlled trials, so the present trial was conducted to definitively test the effectiveness of a custom in-home display to reduce overall electricity consumption among residential customers.

4.1.1. Plausible Mechanism

The literature review included the systematic meta-analysis and additional sources, finding that in-home displays likely reduce electricity consumption by promoting awareness of electricity use, where people realize that they are consuming "an invisible product that is often ignored".[62] The first source of support for this mechanism comes from the finding that over the last forty years more sophisticated in-home displays (e.g., real-time feedback, graphical displays) have not been associated with larger effect sizes. This suggests that it is "the presence of the information itself—not its presentation in a more salient, graphical format—that is causing the behavior change".[63]

[6] http://openwetware.org/wiki/Main_Page
[7] http://www.consort-statement.org/consort-statement/

There is evidence that in-home displays do increase awareness, as Norton *et al.*[64] found that 75% of participants reported being more aware of potential energy efficient actions after interacting with the PowerCost Monitor display. Other evidence shows the association between the in-home display, changes in awareness, and changes in consumption. For example, Hutton *et al.*[65] found the largest effects of the ECI display above and beyond education alone from participants in California, who knew the least about electricity consumption, as opposed to two Canadian cities where participants knew more. Similarly, Yun *et al.*[66] found that participants given an in-home display who had low or moderate awareness of energy consumption at baseline reduced their energy consumption more than those who initially had high awareness. This mechanism is not assured, however, as consciousness of problematic electricity use, and behaviors that contribute to them, is likely to be a necessary but not sufficient condition for behavioral change.[67]

4.1.2. Objectives

The primary hypothesis of the trial was that residential customers who were given in-home displays would reduce their overall electricity consumption more than those given a placebo display. The secondary hypothesis was that participants who report becoming more aware of their electricity use, based on an electricity knowledge test and subjective scale, would benefit from the in-home display, whereas those who did not would not benefit.

4.2. Methods

4.2.1. Trial Design

To test the primary and secondary hypothesis we conducted a sixteen month double-blind randomized concurrent controlled trial comparing an in-home display treatment group against placebo group. No changes were made to methods or eligibility criteria after the trial commenced.

4.2.2. Participants

The sample frame for the study included all participants in a geographic location (e.g., Pennsylvania) who had a smart meter linked to their home or apartment. This inclusion criterion was necessary so that information could be communicated from the smart-meter to the in-home display in real-time. The only other eligibility criterion was that participants must not have expressed that they were likely to move out of their home or apartment during the one year study period.

4.2.3. Interventions

Both the functional in-home display and the placebo provided tips on saving energy, weather, temperature, and date/time capabilities. However, only the functional in-home display provided real-time electricity use information.

4.2.4. Outcomes

Monthly electricity use for each participant was extracted using hourly smart-meter data. The secondary outcome was a self-reported awareness measure taken at both baseline (before randomization) and again at close-out (at the end of the study). Participants were asked to list in an open-ended format all the factors that consume electricity in their household. There were no changes to trial outcomes after the trial commenced.

4.2.5. Sample Size and Power

To estimate statistical power and choose the study sample size, we use the adjusted HLM estimate of the effect size from Davis *et al.*,[1] which had a Cohen's d of 0.63.[68] Based on this expected effect size, a simple two-sample t-test comparison between the average electricity consumption of those in the in-home display versus placebo group, each with 300 participants, would have over 90% power to detect the effect by rejecting the null hypothesis. Even with an effect size 5 times smaller than that suggested by prior evidence, a sample size of 300 in each group would have a 50% chance of detecting the effect. Thus, a sample size of 300 participants in each group provides a study that is very sensitive to effects in the range of plausible values based on prior data. There were no planned interim analyses or stopping guidelines.

4.2.6. Sequence Generation

After 600 participants consented to participate, they were then immediately randomly assigned, using simple randomization, to one of two groups using the following procedure:

1. Every recruited customer in the sample was given a random number from 0 to 1 using the "=rand()" function in Excel 2010.
2. The random numbers were then sorted from smallest to largest.
3. The first 300 customers were assigned group label A.
4. The second 300 customers were assigned group label B.

4.2.7. Allocation Concealment

After random assignment, the customer information and group labels were then recorded on each participant's digital case report form. A neutral third party, who, in addition to the in-home display vendor, was the only

unblinded person, then generated a random number from 0 to 1 using the "=rand()" function in excel for each group label. The group label with the lowest number was assigned to the treatment group, the next lowest to the placebo group. The neutral party then informed the vendor to make two types of "welcome packages." These packages had identical weight, center of gravity, and size, and only differed on whether they contained the active in-home display or placebo control in-home display whether they were labeled A or B. The vendor then mailed the packages to the participants. Both the researchers and the vendor knew only the customer name and the group label of the package that was sent.

4.2.8. Implementation

The random allocation sequence was generated by Jay Apt. Participants were enrolled by Alexander Davis and Tamar Krishnamurti. Participants were assigned to interventions by Alexander Davis and Tamar Krishnamurti.

4.2.9. Blinding

Once the in-home displays were received, the study investigators did not know whether each individual participant had an activated or placebo frame, and all personnel interacting with the participant or the data did not know group assignment. The recruitment document did not provide participants with an expectation of real-time feedback on the in-home display, minimizing the risk of frustrated expectations in the placebo group. The interventions were identical except for the uninformative group label on the outside of the welcome package. Blinding was broken before the manuscript was submitted for publication to correctly label statistical analyses.

4.2.10. Statistical Methods

Simple independent-sample t-tests compared the total electricity used in the in-home display treatment group and placebo group. The mediating factor of awareness was tested by first using simple linear regression of treatment group assignment on awareness, and then treatment group assignment on electricity consumption, controlling for awareness. One additional analysis was performed. The propensity score model was a simple linear regression of treatment group assignment on electricity use controlling for each participant's propensity to volunteer.

4.3. Results

4.3.1. Recruitment

Recruitment began February 1, 2012 and ended April 1, 2012. Follow-up lasted from May, 1 2012 to September 1, 2013. The trial was ended September 1, 2013 as planned.

Based on prior recruitment rates in the medical literature that use recruitment best practices, as well as experiences from other in-home display trials, we expected a 10% recruitment rate. From the set of residential customers who met the inclusion criteria (the sample frame), the sample was selected from the sample frame using the following randomization process:

1. Every customer in the sample frame was given a random number from 0 to 1 using the "=rand()" function in Excel 2010.
2. The random numbers were then sorted from smallest to largest.[8]
3. The first 6000 customers in the sample frame with the lowest random numbers were included in the sample.

Eligible customers included in the sample were mailed a recruitment document that was pre-tested for enrollment quality and constructed based on best practices.[35,69] The recruitment document included information about the study, and provided customers three ways to enroll: by email, by returning an addressed and stamped postcard, or by calling the researchers on a 1-800 number. Recruiters who answered phones or responded to emails were certified for being able to handle participant questions and give accurate information about the study.[70]

Employing an opt-out design, one week after the first recruitment mailing, those who did not respond were sent a post-card reminding them to sign up and eliciting reasons for not responding. Two weeks after the first recruitment all non-responders were contacted by phone to recruit them, inform them about the study, or understand their reasons for refusal. Three weeks after the first recruitment all remaining non-responders were again contacted by phone. Screening logs were maintained to determine reasons for refusal and exclusion to allow us to determine how the enrolled and refused population differ. After three weeks the set of participants who did not opt-out of the study was complete, and if this did not include enough participants then the process was repeated taking another random sample of 6000 customers.

Those who did not opt-out were then contacted by phone to discuss the purpose of the trial, explain the informed consent, and acquaint the participant with the study and requirements. A second call then answered questions raised and reviewed the study requirements a second time. Participants were told, under the requirements of informed consent, that they would be given an in-home display that would provide information along with electricity saving tips. Those assigned to the treatment or placebo in-home display groups

[8]Note that we sorted the values of the rand function and not the rand function itself.

did not know whether they got the treatment in-home display or control in-home display. During the second call the participants were asked to verbally confirm consent to the study, verified by an audio record.

4.3.2. Baseline Data

Demographic characteristics are summarized in Table 1.[9] Participants in the control and treatment group were balanced on both demographic factors.

Table 1: Demographics

Demographic	Treatment	Placebo	t-test (p-value)
Age	43 (15)	43 (14)	0.088 (.93)
Gender (Male)	.48 (.50)	.53 (.50)	1.1 (.25)

4.3.3. Numbers Analyzed

Three hundred participants were initially assigned to each group, and all three hundred from each group were included in all statistical analyses according to their original group assignments.

4.3.4. Outcomes and Estimation

Table 2 shows the average monthly kWh use (standard deviations in parentheses) for the treatment and placebo group for all 16 months in the study. As can be seen, the treatment group had significantly lower average kWh use than the placebo group in every month.

A varying-intercept model was fit allowing each participant and month to have its own intercept.[71] Across the sixteen month period, participants in the treatment IHD group used on average 124 less kWh each month than those in the placebo group $t(598) = 20$, $p < .01$.

4.3.5. Ancillary Analyses

A simple, one-minute questionnaire includes items that were tested for ability to predict the volunteering enrollment decision. Participants were offered a $2 bill[72,73] for completing the questionnaire. The items of the questionnaire were then used to develop a propensity score model to adjust for volunteer bias. The propensity score model adjusted the treatment effect model.[10]

[9] All data, materials, and statistical analyses are available at Harvard's Dataverse http://hdl.handle.net/1902.1/20271 V1 [Version]

[10] We omit the mediation and propensity score model here for brevity, but these would be reported in the actual paper.

Table 2: Average Monthly kWh Use for Treatment IHD and Placebo Group

Month	Treatment	Placebo	t-test (p-value)	Cohen's d
May, 2012	881 (196)	996 (201)	7.1 (.01)	0.58
June, 2012	877 (203)	990 (206)	6.7 (.01)	0.55
July, 2012	865 (206)	1002 (188)	8.5 (.01)	0.69
August, 2012	881 (211)	1024 (213)	8.3 (.01)	0.68
September, 2012	891 (192)	1004 (190)	7.3 (.01)	0.6
October, 2012	862 (205)	998 (196)	8.3 (.01)	0.68
November, 2012	882 (191)	1002 (204)	7.4 (.01)	0.6
December, 2012	878 (209)	989 (197)	6.7 (.01)	0.54
January, 2013	851 (209)	1004 (203)	9.1 (.01)	0.74
February, 2013	864 (194)	987 (202)	7.6 (.01)	0.62
March, 2013	878 (206)	994 (220)	6.7 (.01)	0.55
April, 2013	866 (207)	1007 (194)	8.6 (.01)	0.7
May, 2013	866 (205)	994 (203)	7.7 (.01)	0.63
June, 2013	869 (211)	998 (199)	7.7 (.01)	0.63
July, 2013	876 (219)	994 (210)	6.8 (.01)	0.55
August, 2013	869 (195)	995 (196)	7.9 (.01)	0.64

4.3.6. Harms

Participants were provided with a 1-800 number to register difficulties or complaints. A small number of participants in both the treatment and placebo group reported increased monthly electricity bills due to the addition of the in-home display. No other complaints related to participation in the study were registered.

4.4. Other Information

4.4.1. Registration

The trial was registered with the WHO international Clinical Trials Registry Platform (here).

4.4.2. Protocol

Prior to conducting the trial, the protocol (available here (link)) was peer reviewed, published in the journal Trials.

4.4.3. Funding and Responsibilities

The project involved collaboration between the in-home display vendor, a utility company, and an academic university. The creation of the experimental design, construction of the sampling frame, sampling procedures,

randomization procedures, blinding procedures, as well as statistical analyses and creation of reports were the sole responsibility of the university, as was the responsibility for ensuring that the protocol was administered, amended, and revised appropriately.[11] The creation of the in-home display content is the joint work of the university, utility, and vendor. The creation of communications with participants, including baseline, volunteer, psychodemographic and follow-up surveys were created jointly by the utility and the university. The monitoring and evaluation of the trial was done jointly by the utility, vendor, and university. The vendor covered the cost of the development and production of the in-home displays, whereas the utility covered the logistical costs of the trial, including smart-meters, phone centers, and mailing, and the university covered the costs of employing academic researchers.

5. Costs of a Gold Standard RCT

One primary consideration in conducting a field trial will always be financial constraints. While there is very limited public data on the costs of pilot studies, estimates put the average cost to the utility of an in-home display trial at approximately $500 per-household.[74] Table 3 shows estimates of the increased costs for designing and implementing an in-home display trial that adheres to our gold standard.

Additional costs are primarily equipment-based and do no take into account economies of scale (i.e. equipment may be discounted at higher bulk purchase levels). What we do not present, but is a consideration, is the future cost of having to conduct additional trials as a result of poor evidence collection or the even more considerable cost of rolling out a technology or behavioral intervention that would not have been indicated as effective if evaluated with a better initial pilot study.

6. Conclusion

In this paper we have made the argument that electricity industry decision-makers face decisions about whether to implement new technologies that can affect consumer behavior in the same way that the FDA must make drug or new device regulatory decisions. Just as the FDA requires a high standard of evidence for approval, so too should electricity industry decision-makers require a high standard of evidence before investing in new technologies or behavioral interventions. This standard of evidence is the randomized controlled effectiveness trial. We detailed how these trials are conducted in

[11] http://cancercenters.cancer.gov/

Table 3: Estimated costs for in-home display trial to meet the gold standard. Planning costs assume skilled labor at $50 per hour. Fixed costs only need to be paid one time for the entire study. Marginal costs are costs per participant.

Item	Additional Fixed Costs
Systematic Review	$\sim$ 100-200 hours planning
Sequence Generation	No cost
Allocation Concealment	$\sim$ 10-20 hours planing
Blinding	$\sim$ 10-20 hours planning
Eligibility Exclusions	$\sim$ 30-50 hours planning
Volunteer Adjustment	$\sim$ 50-100 hours planning
Withdrawal Prevention	$\sim$ 50-100 hours planning
Reporting	$\sim$ 10-20 hours labor
Sum	$\sim$ \$13,000-\$26,000 (260-510 labor hours)

Item	Additional Marginal Costs
Placebo Control Group	$\sim 2\times$ as many participants
Volunteer Adjustment	$\sim$ \$5-\$10 per participant
Withdrawal Prevention	$\sim$ \$5-\$10 per participant
Sum	$\sim$ \$510-\$520 per treatment-control pair

biomedical research, and have provided an example for how they can be translated to research on electricity consumption behavior, using in-home displays as an example. Finally, we provide two checklists in Appendix A to help researchers implement these standards, as checklists have been shown to be very effective in implementing well-known procedures.[75,76] We hope that those who conduct research on human behavior in electricity learn and use these standards, and that before investments are made, policy-makers demand the high level of evidence that their customers, constituents, the public, and scientific community deserve.

Appendix A. Checklists

Table A.4: Background, Internal Validity, and External Validity

Item	Yes	No
Background		
Has a systematic review been completed?		
Has a meta-analysis been completed?		
Has the methodological quality of prior evidence been accounted for?		
Have objectives been clearly stated as hypotheses?		
Have plausible mechanisms been identified and facilitated by design?		
Internal Validity		
Has a concurrent control group been used?		
Is the control group identical to the treatment group?		
Is the placebo indistinguishable from the treatment?		
Has equipoise been established?		
Is randomization used?		
Is the randomizing sequence truly random?		
Is random allocation adequately concealed?		
Are participants blinded to their group?		
Are personnel, data collectors, and data analysts blinded?		
Are contacts and visits balanced to maintain blinding?		
External Validity		
Is the sample drawn randomly from the population?		
Are the eligibility criteria clear, pre-specified, and minimal?		
Are justifications for exclusion criteria provided?		
Is a propensity score model used to account for volunteering?		
Do exclusion criteria minimize chances of withdrawal?		
Are measurements likely to be completed before withdrawal?		

Table A.5: Statistical Validity and Reporting

Item	Yes	No
Statistical Validity		
Are data analysts blinded?		
Are statistical analyses preplanned and published prior to study commencement?		
Are pre-planned and post-hoc analyses clearly identified?		
Has a power analysis been conducted?		
Has a sample size calculation been conducted?		
Are the rules for early stopping set forth ahead of time?		
Has a whole-sample analyses been done?		
Has intention-to treat been used?		
Has imputation been used?		
Reporting		
Does the trial follow the CONSORT statement?		
Are data made publicly available?		
Are statistical analyses publicly available and reproducible?		
Are study materials publicly available and reproducible?		

Appendix B. Acknowledgement

This work was supported by the center for Climate and Energy Decision Making (SES-0949710), through a cooperative agreement between the National Science Foundation and Carnegie Mellon University.

This material is based upon work supported by the Department of Energy under Award Numbers DE-OE0000300 and DE-OE0000204. Disclaimer: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights, Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Endnotes:

[1] A. Davis, T. Krishnamurti, B. Fischhoff, and W. Bruine de Bruin. Setting a standard for electricity field studies. *Unpublished Manuscript*, 2012.

[2] A. Faruqui, S. Sergici, and A. Sharif. The impact of informational feedback on energy consumption–a survey of the experimental evidence. *Energy*, 35(4):1598–1608, 2010.

[3] S.R. Tunis, D.B. Stryer, and C.M. Clancy. Practical clinical trials. *JAMA: the Journal of the American Medical Association*, 290(12):1624–1632, 2003.

[4] L.M. Friedman, C.D. Furberg, and D.L. DeMets. *Fundamentals of clinical trials*. Springer, 2010.

[5] C.L. Meinert. *ClinicalTrials: Design, Conduct and Analysis*. OUP USA, 2012.

[6] G.H. Guyatt, A.D. Oxman, G.E. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coello, and H.J. Schünemann. Grade: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*, 336(7650):924–926, 2008.

[7] M. Roland and D.J. Torgerson. Understanding controlled trials: What are pragmatic trials? *British Medical Journal*, 316(7127):285, 1998.

[8] D.C. Cherkin, D. Eisenberg, K.J. Sherman, W. Barlow, T.J. Kaptchuk, J. Street, and R.A. Deyo. Randomized trial comparing traditional chinese medical acupuncture, therapeutic massage, and self-care education for chronic low back pain. *Archives of Internal Medicine*, 161(8):1081–1088, 2001.

[9] R.M. Turner, D.J. Spiegelhalter, G. Smith, and S.G. Thompson. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):21–47, 2009.

[10] R. Scheines. The similarity of causal inference in experimental and non-experimental studies. *Philosophy of Science*, 72(5):927–940, 2005.

[11] J. Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2000.

[12] P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, prediction, and search*. The MIT Press, 2000.

[13] R. Kunz and A.D. Oxman. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal*, 317(7167):1185–1190, 1998.

[14] D.P. Byar, R.M. Simon, W.T. Friedewald, J.J. Schlesselman, D.L. DeMets, J.H. Ellenberg, M.H. Gail, J.H. Ware, et al. Randomized clinical trials. perspectives on some recent ideas. *The New England Journal of Medicine*, 295(2):74–80, 1976.

[15] DG Altman and CJ Dore. Randomisation and baseline comparisons in clinical trials. *The Lancet*, 335(8682):149–153, 1990.

[16] K.F. Schulz and D.A. Grimes. Generation of allocation sequences in randomised trials: chance, not choice. *The Lancet*, 359(9305):515–519, 2002.

[17] G.S. Tune. Response preferences: A review of some relevant literature. *Psychological Bulletin*, 61(4):286–302, 1964.

[18] K.F. Schulz, D.A. Grimes, et al. Allocation concealment in randomised trials: defending against deciphering. *The Lancet*, 359(9306):614–618, 2002.

[19] K.F. Schulz et al. Randomised trials, human nature, and reporting guidelines. *The Lancet*, 348(9027):596–598, 1996.

[20] KF Schulz. Unbiased research and the human spirit: the challenges of randomized controlled trials. *CMAJ: Canadian Medical Association Journal*, 153(6):783–786, 1995.

[21] I. Chalmers. Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments. *International Journal of Epidemiology*, 30(5):1156–1164, 2001.

[22] M.T. Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11):776–783, 1962.

[23] B. Franklin, Majault, L. Roy, Sallin, J.S. Bailly, D'arcet, D. Bory, J.I. Guillotin, and A. Lavoisier. Report of the commissioners charged by the

king with the examination of animal magnetism. *International Journal of Clinical and Experimental Hypnosis*, 50(4):332–363, 2002.

[24] M.T. Haahr and A. Hróbjartsson. Who is blinded in randomized clinical trials? *The Cochrane Collaboration Methods Groups Newsletter*, 3:14, 2006.

[25] L.E. Miller and M.E. Stewart. The blind leading the blind: use and misuse of blinding in randomized controlled trials. *Contemporary Clinical Trials*, 32(2):240–243, 2011.

[26] K.F. Schulz, D.A. Grimes, D.G. Altman, and R.J. Hayes. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *British Medical Journal*, 312(7033):742–744, 1996.

[27] D.L. Sackett. Commentary: Measuring the success of blinding in rcts: don't, must, can't or needn't? *International Journal of Epidemiology*, 36(3):664–665, 2007.

[28] G.H. Guyatt, D.L. Sackett, D.J. Cook, Evidence-Based Medicine Working Group, et al. Users' guides to the medical literature. ii. how to use an article about therapy or prevention. b. what were the results and will they help me in caring for my patients. *JAMA: the Journal of the American Medical Association*, 271(1):59–63, 1994.

[29] H.G.C. Van Spall, A. Toren, A. Kiss, and R.A. Fowler. Eligibility criteria of randomized controlled trials published in high-impact general medical journals. *JAMA: the Journal of the American Medical Association*, 297(11):1233–1240, 2007.

[30] P. Smith and H. Arnesen. Mortality in non-consenters in a post-myocardial infarction trial. *Journal of Internal Medicine*, 228(3):253–256, 1990.

[31] R. Rosenthal and R.L. Rosnow. *The Volunteer Subject.* John Wiley & Sons, 1975.

[32] J.R. Wright, D. Crooks, P.M. Ellis, D. Mings, and T.J. Whelan. Factors that influence the recruitment of patients to phase III studies in oncology. *Cancer*, 95(7):1584–1591, 2002.

[33] EG Bevan, LC Chee, SM McGhee, and GT McInnes. Patients' attitudes to participation in clinical trials. *British Journal of Clinical Pharmacology*, 35(2):204–207, 1993.

[34] K. Cox and J. McGarry. Why patients donât take part in cancer clinical trials: an overview of the literature. *European Journal of Cancer Care*, 12(2):114–122, 2003.

[35] S. Treweek, M. Pitkethly, J. Cook, M. Kjeldstrøm, T. Taskila, M. Johansen, F. Sullivan, S. Wilson, C. Jackson, R. Jones, and E. Mitchell. Strategies to improve recruitment to randomised controlled trials. *Cochrane Database of Systematic Reviews*, 4, 2010.

[36] B. Williams, L. Irvine, A.R. McGinnis, ME McMurdo, I.K. Crombie, et al. When "no" might not quite mean "no"; the importance of informed and meaningful non-consent: results from a survey of individuals refusing participation in a health-related research project. *BMC Health Serv Res*, 7:59–69, 2007.

[37] C. Junghans, G. Feder, H. Hemingway, A. Timmis, and M. Jones. Recruiting patients to medical research: double blind randomised trial of "opt-in" versus "opt-out" strategies. *British Medical Journal*, 331(7522):940–948, 2005.

[38] J.M. Wooldridge. *Introductory Econometrics: A Modern Approach*. South-Western Pub, 2009.

[39] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press New York, 2007.

[40] Y.H. Koog, M. Gil, S.R. We, H. Wi, and B.I. Min. Barriers to participant retention in knee osteoarthritis clinical trials: A systematic review. In *Seminars in Arthritis and Rheumatism*. Elsevier, 2012.

[41] M.S. Jordhøy, S. Kaasa, P. Fayers, G. Underland, and M. Ahlner-Elmqvist. Challenges in palliative care research; recruitment, attrition and compliance: experience from a randomized controlled trial. *Palliative Medicine*, 13(4):299–310, 1999.

[42] K.A. Robinson, C.R. Dennison, D.M. Wayman, P.J. Pronovost, and D.M. Needham. Systematic review identifies number of strategies important for retaining study participants. *Journal of Clinical Epidemiology*, 60(8):757–765, 2007.

[43] S. Hollis and F. Campbell. What is meant by intention to treat analysis? survey of published randomised controlled trials. *British Medical Journal*, 319(7211):670–674, 1999.

[44] J.G. Ibrahim, M.H. Chen, S.R. Lipsitz, and A.H. Herring. Missing-data methods for generalized linear models. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

[45] S. Al-Marzouki, I. Roberts, T. Marshall, and S. Evans. The effect of scientific misconduct on the results of clinical trials: a delphi survey. *Contemporary Clinical Trials*, 26(3):331–337, 2005.

[46] E.E. Leamer. Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43, 1983.

[47] B. Fischhoff. *For those condemned to study the past: Heuristics and biases in hindsight*, pages 335–351. New York: Cambridge University Press, 1982.

[48] J.P. Simmons, L.D. Nelson, and U. Simonsohn. False-positive psychology. *Psychological Science*, 22(11):1359–1366, 2011.

[49] J.A. Freiman, T.C. Chalmers, H. Smith Jr, and R.R. Kuebler. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *New England Journal of Medicine*, 299(13):690–694, 1978.

[50] J. Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992.

[51] R. Simon. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, 10(1):1–10, 1989.

[52] P.C. Gøtzsche. Blinding during data analysis and writing of manuscripts. *Controlled Clinical Trials*, 17(4):285–290, 1996.

[53] K. Schulz, D. Altman, D. Moher, et al. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, 8(1):8–18, 2010.

[54] D. Moher, S. Hopewell, K.F. Schulz, V. Montori, P.C. Gøtzsche, PJ Devereaux, D. Elbourne, M. Egger, D.G. Altman, D. Moher, et al. Consort 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *British Medical Journal*, 340:c869, 2010.

[55] E.H. Turner, A.M. Matthews, E. Linardatos, R.A. Tell, and R. Rosenthal. Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, 358(3):252–260, 2008.

[56] R. De Vries, M.S. Anderson, and B.C. Martinson. Normal misbehavior: Scientists talk about the ethics of research. *Journal of Empirical Research on Human Research Ethics: JERHRE*, 1(1):43–50, 2006.

[57] B.C. Martinson, M.S. Anderson, and R. De Vries. Scientists behaving badly. *Nature*, 435(7043):737–738, 2005.

[58] B. Goldacre. Bad pharma. *Fourth Estate*, 2012.

[59] E. Yong. Bad copy. *Nature*, 485(7398):298–300, 2012.

[60] V. Stodden. Enabling reproducible research: Open licensing for scientific innovation. *International Journal of Communications Law and Policy, Forthcoming*, 2009.

[61] F. Leisch. Sweave. dynamic generation of statistical reports using literate data analysis. In *COMPSTAT 2002–Proceedings in Computational Statistics*. Vienna University of Economics and Business, 2002.

[62] J. Schembri. The influence of home energy management systems on the behaviours of residential electricity consumers: An ontario, canada case study. 2008.

[63] D. Allen and K. Janda. The effects of household characteristics and energy use consciousness on the effectiveness of real-time energy use feedback: a pilot study. In *Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings*, pages 7.1–7.12, 2006.

[64] Norton. Powercost monitor pilot program evaluation. Technical report, Blueline, 2008.

[65] R.B. Hutton, G.A. Mauser, P. Filiatrault, and O.T. Ahtola. Effects of cost-related feedback on consumer knowledge and consumption behavior: A field experimental approach. *Journal of Consumer Research*, pages 327–336, 1986.

[66] T.J. Yun. Investigating the impact of a minimalist in-home energy consumption display. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pages 4417–4422. ACM, 2009.

[67] C. Fischer. Feedback on household electricity consumption: a tool for saving energy? *Energy Efficiency*, 1(1):79–104, 2008.

[68] J. Cohen. The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65(3):145–153, 1962.

[69] PJ Edwards, I. Roberts, MJ Clarke, C. DiGuiseppi, R. Wentz, I. Kwan, R. Cooper, LM Felix, and S. Pratap. Methods to increase response to postal and electronic questionnaires. *The Cochrane Library*, 2009.

[70] J.L. Donovan, J. Athene Lane, T.J. Peters, L. Brindle, E. Salter, D. Gillatt, P. Powell, P. Bollina, D.E. Neal, and F.C. Hamdy. Development of a complex intervention improved randomization and informed consent in a randomized controlled trial. *Journal of Clinical Epidemiology*, 62(1):29–36, 2009.

[71] A. Gelman, Y.S. Su, M. Yajima, J. Hill, M.G. Pittau, J. Kerman, and T. Zheng. arm: Data analysis using regression and multilevel/hierarchical models. *R package version*, pages 1–3, 2010.

[72] M.M. Doody, A.S. Sigurdson, D. Kampa, K. Chimes, B.H. Alexander, E. Ron, R.E. Tarone, and M.S. Linet. Randomized trial of financial incentives and delivery methods for improving response to a mailed questionnaire. *American Journal of Epidemiology*, 157(7):643–651, 2003.

[73] C.L. Booker, S. Harding, and M. Benzeval. A systematic review of the effect of retention methods in population-based cohort studies. *BMC Public Health*, 11(1):249–261, 2011.

[74] B. Foster and S. Mazur-Stommen. Results from recent real-time feedback studies. Technical report, American Council for an Energy-Efficient Economy, 2012.

[75] A. Gawande. *The Checklist Manifesto: How to Get Things Right*. Profile Books, 2010.

[76] P. Pronovost, D. Needham, S. Berenholtz, D. Sinopoli, H. Chu, S. Cosgrove, B. Sexton, R. Hyzy, R. Welsh, G. Roth, et al. An intervention to decrease catheter-related bloodstream infections in the icu. *New England Journal of Medicine*, 355(26):2725–2732, 2006.