Setting a standard for electricity pilot studies

Alexander Davis [a,b]

Tamar Krishnamurti [c,d]

Baruch Fischhoff [a,d]

Wandi Bruine de Bruin [a,d]

Carnegie Mellon University

Pittsburgh, PA 15213

June 7, 2012

[a]Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213.

[b]To whom correspondence should be addressed. Email: alexander.l.davis1@gmail.com; Phone: 412-268-1207; Fax: 412-268-6938. http://dvn.iq.harvard.edu/dvn/dv/alexdavis

[c]Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213. Email: tamar@cmu.edu

[d]Department of Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA 15213. Email: baruch@cmu.edu

Author Contributions: A.D., T.K., B.F., and WBB designed research and wrote the paper; A.D. and T.K. performed research; A.D. developed risk-of-bias meta-analysis and analyzed the data.

Abstract

In-home displays, dynamic pricing, and automated devices aim to reduce residential electricity use – overall and during peak hours. We present a meta-analysis of 32 studies examining the impacts of these interventions. We find that methodological problems were common in the design of these studies, leading to artifactually inflated results relative to what one would expect were these interventions implemented in the general population. Particular problems included having volunteer participants who may have been especially motivated to reduce their electricity use, letting participants chose their preferred intervention, and having high participant attrition rates. Using bias adjustment factors from medical clinical trials, we calculate effectiveness estimates less than half those reported in the reviewed studies. Our analyses are limited by the incomplete reporting of many studies. Within that constraint, we find that in-home displays were the most effective intervention for reducing overall electricity use (~4% using reported data; ~3% after adjusting for bias), while dynamic pricing significantly reduced peak demand (~12% with reported data; ~6% after adjusting), especially with home automation (~24% with reported data; ~13% after adjusting). We conclude with recommendations for designing and reporting evaluation studies, so as to improve the return on the resources invested in them.

Word count: 198

Classifications: Demand-side management; Electricity end-use; Energy Use Behavior; Environmental Management Systems (EMS);

Keywords: behavior; conservation; in-home-display;

1. Introduction

Reducing overall residential electricity use will lower emissions of greenhouse gases and other pollutants (Weisser, 2007) and decrease the need for additional power plants and transmission capacity (FERC, 2009). Reducing residential electricity use during peak demand times (e.g., hot summer afternoons) will lower the risk of blackouts and the need for back-up facilities. Currently, 15% of generation and transmission capacity in the Mid-Atlantic States is used less than 1% of the time (Spees and Lave, 2007). As a result, there have been many studies of interventions designed to reduce that waste.

The three most common interventions are (a) *in-home displays* that provide feedback about energy consumption; (b) *dynamic pricing* programs where residential electricity prices follow the wholesale market, creating an incentive to reduce use during peak-demand hours; and (c) *automation*, with programmable thermostats, smart switches, and other technologies.

Although many studies have evaluated the effectiveness of such interventions, their experimental designs and reporting protocols vary so much that it is hard to aggregate their results. Here, we propose and apply a standard approach, based on the *risk-of-bias* (RoB) methodology developed to improve medical clinical trials (Higgins, Altman and Sterne, 2011; Moher, Hopewell, Schulz *et al.*, 2010) and applied to treatments as diverse as asthma (Hartling, Bond, Vandermeer, *et al.*, 2011), routine antenatal care (Turner, Spiegelhalter, Smith, *et al.*, 2009), and influenza treatment and prevention (Shun-Shin, Thompson, Heneghan, *et al*., 2009).

Risk-of-bias analysis extends methodological standards for medical research (http://www.cochrane.org/ and http://www.consort-statement.org/; Moher, Hopewell, Schulz, *et al*., 2010) by accommodating the finding that flawed studies often found positive health effects that vanished, or reversed, with sounder ones (Moher, Pham, Jones, *et al*. 1998). For example, the initial promise of hormone replacement therapy (Petitti, 2004) was later found to reflect *selection bias*, whereby women who opted to take the therapy had relatively high socio-economic status, which is associated with better health outcomes (Grady, Herrington, Bittner, *et al*., 2002). The risk-of-bias approach adjusts reported effect sizes for the impacts of the most common biases in medical research. We apply it here to studies estimating the effects of interventions on residential electricity use, first by identifying those biases and then by adjusting reported treatment effects, using correction factors for medical trials. Our approach also provides guidance for designing and reporting field trials.

Reviews of studies evaluating interventions targeting residential electricity use often note problems in their design (Abrahamse, Steg, Vlek, *et al*., 2005; Carrol, Hatton, and Brown, 2009; Darby, 2001; Darby, 2006; Ehrhardt-Martinez, Donnely, and Liatner, 2010; Faruqui, Hledik and Sergici, 2009; Faruqui, Sergici and Sharif, 2010; Fischer, 2008; IEA, 2007; Neenan, 2009; Roberts and Baker, 2003) and reporting (Fischer, 2008). Indeed, Abrahamse, Steg, Vlek *et al.* (2005) concluded that reporting was so deficient that a thorough meta-analysis was infeasible. Our review considers both reporting

practices and the biases revealed by those details that are provided.

In its systematic reviews of medical studies, the Cochrane Collaboration has identified five common biases with serious effects (Higgins, Altman, and Sterne, 2011): 1) *selection* bias, arising when participants decide whether and how to participate; 2) *attrition* bias, arising when participants are excluded or withdraw from a study; 3) *performance* bias, arising when experimenters know participants' group assignment; 4) *detection* bias, arising when researchers' knowledge of group assignment affects their interpretation of participants' behavior; and 5) *reporting* bias, arising when researchers omit details in their reports. The following sections briefly discuss the threats that these biases may pose to electricity field trials.

Selection biases arise when people who receive an intervention differ from those who do not – limiting researchers' ability to establish causality and generalize study results to the general population. One variant, *intervention selection bias,* occurs when participants choose their treatment group, rather than being randomly assigned (Altman and Bland, 1999). Stukel, Fisher, Wennberg, *et al.* (2007) estimated that studies with this bias reported 44% greater treatment effects than ones that using a predetermined assignment process. If people who select a treatment are especially motivated to change their behavior, this bias leads to overestimating intervention effectiveness. For example, the Olympic Peninsula Pilot Hammerstrom, Ambrosio, Brous, *et al*. (2007) randomly assigned participants to the control or intervention condition, but allowed those receiving an intervention to choose among three pricing

options: fixed rate, time of use with critical peak pricing, and real-time pricing. These results could not be confidently generalized to universal adoption of any of these options, if participants chose the plan best suited to them (e.g., if people who could, or would not shift their electricity use to off-peak hours chose the fixed-rate plan, while those who could and would make that shift selected other plans). Generalization would also be limited if people chose plans for reasons unrelated to their incentives. For example, if people with less education (and less income) choose a fixed rate program because it is simpler and easier to understand, its impact will be underestimated because their lower initial electricity consumption offers less room for reductions.

A second version of this bias is *volunteer selection bias*, arising when participants are recruited through advertisements. Volunteers differ from randomly sampled individuals in many ways that might bias treatment effects (Callahan, Hojat and Gonella, 2007; Rosenthal and Rosnow, 1975; Barclay, Todd, Finlay, *et al.*, 2002). For example, Sulyma, Tiedemann, Pedersen *et al*. (2008) found higher education and income levels among volunteer British Columbia Hydro customers. If, as a result, they are also better able to comprehend and respond to program information, then studies involving them will overestimate programs' general effectiveness. Baladi, Herriges, and Sweeney (1998) found that volunteers were better than non-volunteers at estimating peak-demand electricity use in a time-of-use experiment and more optimistic about program benefits. Generally speaking, one cannot assume that being similar in some respects (e.g., demographics, baseline electricity usage)

means responding similarly to treatments (Train, McFadden and Goett, 1987)). (Note that one could generalize from a study with volunteer bias to an actual program that recruited participants in the same way – e.g. by enrolling relatively wealthy, well-educated, optimistic consumers.)

A third form of selection bias, *sequence generation bias,* occurs when participants are assigned to interventions by a non-random process, such as alphabetically or by alternating assignments. For example, the Baltimore Gas and Electric (Faruqui and Sergici, 2009a) Smart Energy Pricing Pilot, recruited consumers for dynamic pricing first, then for peak-time rebate, and so on. If people who are most eager to receive some intervention sign up first, serial enrollment will overestimate the effectiveness of the first intervention. Schulz (Schulz, 1995a) found larger treatment effects in clinical trials with inadequate randomization.

Formal randomization procedures are needed because people cannot generate random sequences on their own (Tune, 1964; Tversky and Kahneman, 1971) and may be tempted to make assignments favorable to their interests or to let participants do so (e.g., postponing medical treatment until a clinical trial begins). Schulz and Grimes (2002) found that 5% of clinical trials reporting random assignment actually used deterministic rules (e.g., alternation, date of birth, day of hospital admission). An additional 63% reported too few details to determine their method.

A fourth selection bias, *allocation concealment bias,* occurs when

participants or researchers know the assignment sequence (even if random) and can manipulate assignment. Schulz (1995b) and Schulz and Grimes (2002) report researchers trying to decipher allocation sequences, for example, by holding assignment envelopes up to a light (Carleton, Sanders and Burack, 1960; Jüni, Altman and Egger, 2001). In an electricity field study, such a bias might involve favoring larger houses for home automation, thinking that they will benefit the most or responding to pressure from their owners.

Clinical researchers have long known how knowledge of condition can affect the behavior of experimenters (e.g., giving better care to patients receiving a treatment) or participants (e.g., feeling neglected in a control group). Hutton, Mauser, Filiatrault, and Ahtola (Hutton, Mauser, Filiatrault, *et al.,* 1986) found that consumers told that they were in an energy consumption study used less electricity than customers who were not told (310 kWh vs. 270 kWh), even though neither group received an actual intervention. That difference might reflect a Hawthorne effect, where just knowing one is being studied changes behavior (Parsons, 1974; Orne, 1962).

As an example of the care needed to conceal conditions, Karlowski, Chalmers, Frenkel *et al.* (1975) found that some participants in a study on the effects of ascorbic acid on the common cold guessed their condition based on the taste of their pills. Those in the placebo group who guessed their condition reported more severe symptoms and were more likely to drop out, compared to those who thought that they had received the intervention. Correcting for these problems eliminates what little evidence there was for a protective effect of

ascorbic acid.

Attrition bias occurs when participants are excluded or withdraw from a study for reasons related to the assigned intervention. A field trial's effect will be overestimated if people who see no benefit are more likely to withdraw, leaving no record of their lack of change. The same (or the opposite) could be true for people who drop out because they move, pass away, are hard to reach, have more problems, or fall out with researchers. Even when conditions have equal attrition rates, the causes may be different and bias the results. Faruqui and Sergici (2009b) report a 1% monthly rate of people leaving one field trial, for reasons unknown.

2. Results

As seen in Table 1, all studies overstate the mean and understate the variance. Both adjustments would be larger, if the literature provided an estimate for volunteer bias, which tends to increase treatment effects and reduce treatment variance.

[Insert Table 1]

Biases	Mean Adjustment	Variance Adjustment	Overall Usage (N)	Peak Usage (N)
4,6	1.20	0.017	3	3
3,4,5,6	1.63	0.217	10	6
3,4,5	1.76	0.235	2	3
2,3,4	2.06	0.424	23	37
2,3,4,5,6	2.35	0.630	17	8
2,3,4,5	2.54	0.695	2	0

Volunteer = 1, Intervention = 2, Sequence Generation = 3, Allocation Concealment = 4, Blinding = 5, Attrition = 6

We used several statistical procedures to aggregate results across the 32 studies. Figure 1 shows analyses for the three most widely studied peak usage interventions, aggregated with the Generic Inverse Variance (GIV) method, which weights studies by the inverse of their within-group variance (for the minority of studies reporting it) (Higgins, Altman and Sterne, 2011) (see Appendix B for more details). The top half shows the interventions' effects on residential peak use reported in the study, with no adjustment for bias. For example, the studies reported that Dynamic Pricing and Automation reduced peak consumption by 32.8%, on average, with a 95% confidence interval of approximately 22% to 44%. The lower half of the table adjusts these estimates for risk of bias. It reveals intervention effects that are about half as large and confidence intervals about twice as large. All three confidence intervals now include the possibility of no effect – although the means all indicate some positive effect. These analyses used the Review Manager software (Review Manager, 2011).

[Insert Figure 1]

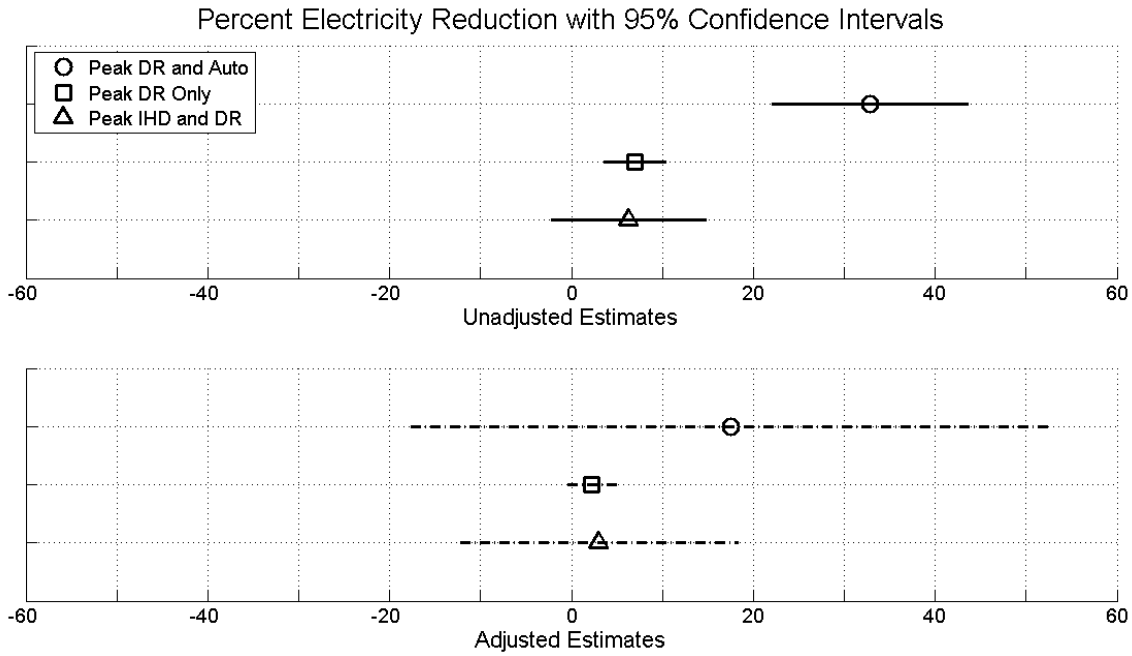

Table 2 summarizes the analyses for all interventions targeting peak-demand residential electricity use, with two aggregation methods: GIV and Hierarchical Linear Models (HLM), which considers only the variance across group means, hence can be calculated for studies with incomplete reporting. GIV gives more weight to studies with less variance, assuming that have better methods, with more consistently delivered interventions. With both aggregation methods, using Dynamic Pricing alone produces significant reductions, which are roughly halved by the risk-of-bias adjustment. Adding Automation substantially increases those effects. Adding In-home displays contributes little to Dynamic Pricing either alone or combined with Automation. Aggregation with Ordinary Least Squares (OLS), which also ignores within-group variance, reveals similar patterns (not shown).

[Insert Table 2]

	Generic Inverse Variance			Hierarchical Linear Model		
Intervention	Un-adjusted Mean (SE)	Adjusted Mean (SE)	Studies (Missing)	Un-adjusted Mean (SE)	Adjusted Mean (SE)	Studies (Missing)
In-home display only	NA	NA	NA	NA	NA	NA
Dynamic Pricing only	**6.93*** (1.74)	**2.49*** (0.64)	11 (17)	**11.56*** (2.26)	**5.92*** (1.21)	28 (3)
In-Home Display with Dynamic Pricing	6.30 (4.34)	3.06 (4.44)	1 (10)	**14.25*** (3.56)	**7.35*** (1.88)	10 (1)
Dynamic Pricing with Automation	**32.80*** (5.50)	**12.53*** (4.64)	3 (16)	**24.40*** (2.86)	**12.68*** (1.52)	16 (3)
In-Home Display with Dynamic Pricing and Automation	NA	NA	0 (4)	**25.78*** (6.30)	**12.56*** (3.30)	3 (1)

**p<0.05*

Table 3 reports comparable statistics for overall reductions in residential electricity use for the five interventions reported in any of the 32 studies. GIV (calculated for studies reporting within-group variances) found the greatest reduction when households received in-home displays alone, equal to 4% in the reported data and 1.2% after adjustment for risk of bias. With HLM, in-home displays alone were again most effective, reducing usage by about 5% in the reported data and 3% after bias adjustment. OLS produced similar results.

Although all five interventions show lower overall usage, after risk-of-bias adjustment, that difference is statistically significant only for in-home displays.

[Insert Table 3]

	Generic Inverse Variance			Hierarchical Linear Model		
Intervention	Un-adjusted Mean (SE)	Adjusted Mean (SE)	Studies (Missing)	Un-adjusted Mean (SE)	Adjusted Mean (SE)	Studies (Missing)
In-home display only	**4.00*** (1.33)	1.20 (1.10)	12 (5)	**5.10*** (2.33)	**2.99*** (1.20)	16 (0)
Dynamic Pricing only	**2.83*** (0.97)	0.30 (0.35)	10 (18)	1.73 (1.17)	0.91 (0.61)	21 (10)
In-Home Display with Dynamic Pricing	2.17 (3.57)	1.05 (2.94)	1 (9)	2.23 (1.41)	1.32 (0.87)	8 (3)
Dynamic Pricing with Automation	**3.61*** (1.58)	0.09 (0.14)	4 (14)	3.84 (2.21)	3.19 (1.92)	11 (7)
In-Home Display with Dynamic Pricing and Automation	**3.00*** (0.323)	1.28 (2.18)	1 (3)	3.00	1.28	1 (3)

** p<0.05*

3. Discussion

We report a critical meta-analysis of all 32 publicly available evaluation studies estimating the effectiveness of interventions designed to reduce overall or peak residential electricity use in the US or Canada: In-Home Displays (IHDs), Dynamic Pricing, and Automation (through programmable thermostats and smart switches). Using the risk-of-bias approach (RoB) developed for medical clinical trials, we found that most studies had methodological features expected to inflate the effectiveness of their focal interventions, relative to routine use in the general population. For example, 27 of the 32 studies used volunteers; 20 allowed researchers or participants to select their interventions rather than using random assignment. When studies reported enough detail to assess their vulnerability to bias, we used estimates from RoB research to

adjust their means and variance. We aggregated the estimates from individual studies using three metaanalytic procedures, Generalized Inverse Variance (GIV), Hierarchical Linear Models (HLM), and Ordinary Least Squares (OLS), which produce similar patterns but somewhat different estimates of effect size.

For peak energy use, the adjustment roughly halved the observed effect size (Table 2). Using reported or adjusted estimates and each aggregation procedure, Dynamic Pricing produced statistically significant savings. Those increased markedly with Automation, but not with IHDs – which also added nothing to the combination of Dynamic Pricing and Automation. Thus, Automation helps, whereas asking consumers to monitor a display does not, even with the incentives offered by Dynamic Pricing. Unfortunately, no study examined the effects of Automation or IHDs alone.

For overall energy usage (Table 3), most interventions showed statistically significant reductions using the reported data and the more sensitive GIV aggregation method (where studies reporting sufficient detail to apply it). The exception was Dynamic Pricing with IHDs. RoB adjustments reduced all mean effects to 3% or less, leaving one statistically significant reduction: In-Home Displays alone, when aggregated with HLM. Only Dynamic Pricing, which significantly reduced peak demand, had weak effects on overall demand. It even seemed to reduce the usefulness of IHDs, a combination that was no more effective than Dynamic Pricing alone with peak demand, perhaps reflecting cognitive overload on consumers.

Almost none of the 32 studies reported enough information to assess its

vulnerability to sequence generation, allocation concealment, and blinding bias. Many lacked detail relevant to intervention or volunteer selection bias. Although that lack of information might mean that having appropriate methods went without saying, RoB analyses based on fuller reporting might well yield even smaller effects. Without within-group variances, we could not apply the more sensitive GIV aggregation method to many studies. It appears that reporting standards have yet to emerge in the gray literature where most of these reports are found. Without them, researchers and practitioners cannot take full advantage of the observations from these trials.

Without risk-of-bias adjustments, many interventions seem to reduce overall usage; with them, only IHDs d (by 1.2%). For peak usage, Dynamic Pricing, especially when combined with Automation, is effective either way, although the effect size is halved with RoB adjustments. However, our bias-adjustment estimates are based on studies of medical clinical trials. There is a vital need for studies examining how far these estimates can be generalized to electricity field trials. The same mechanisms seem plausible (e.g., allocating treatments to people who seem most likely to benefit from them). However, there is no substitute for evidence.

Our conclusions are limited to existing studies, which lacked the combinations of interventions needed to clarify the joint and separate effects of the three forms of intervention. As a result, we cannot, for example, untangle just how IHDs help and hinder performance, when used with other interventions, or how well Automation does by itself. Although full factorial

designs may be prohibitively expensive, fractional factorial designs or response surface methods can provide good alternatives (Myers, Montgomery and Anderson-Cook, 2009), allowing efficient use of limited resources.

Our conclusions are also limited by differences among the studies for each class of interventions, whose members might have varied in their effectiveness (e.g., Dynamic Pricing programs with different incentives). For example, some studies provided consumers with detailed information about their electricity use, while others sent regular monthly bills and yet others required customers to visit websites for their feedback; the studies explained their interventions in different ways, likely varying in how well customers understood them and saw their opportunities to benefit. Studies varying in so many ways beyond the interventions being studies are *meta-confounded* (Deeks, Dinnes, D'Amico, *et al*. 2003).

Although we used three different procedures to aggregate results (GIV, HLM, OLS), producing generally similar results, other approaches would be possible, especially if within-group variances were routinely reported (Ades and Sutton, 2006; Greenland, 2005; Ioannidis, 2011; Spiegelhalter and Best, 2003; Wolpert and Mengersen, 2004).

A straightforward way to improve the reporting, and value of such studies is to adopt the widely used CONSORT guidelines for medical clinical trials (http://www.consort-statement.org/home/), including open web access to supplementary materials providing enough detail to allow replicating each study. Doing so will, among other things, provide the within-group variances.

In addition, we have two recommendations specific to the reporting of electricity field studies: (a) Report effects on both overall and peak usage, in order to see how interventions with one goal affect the other. (b) Report full usage statistics for all time periods for all intervention groups.

We make the following recommendations to avoid the six biases studied here. Although we recognize that practical concerns (e.g., Public Utility Commission approval) may constrain investigators, there are studies that have addressed each threat.

Intervention selection bias can be avoided by randomizing consumers to interventions *after they have agreed to enroll in the study*. For example, the Iowa Residential Electricity Study (RES) randomly assigned participants to time-of-use pricing or control groups after they had signed up to participate (Baladi, Herriges and Sweeney, 1998). See also BC Hydro's AMI study (Sulyma, Tiedemann, Pedersen *et al.*, 2008), the Twin Rivers study (Seligman, Darley and Becker, 1978), and AmerenUE's Residential TOU pilot (Puckett *et al,* 2004).

Sequence generation bias can be avoided by using formal randomization or statistical corrections that adjust estimated effects based on observed factors (e.g., propensity score matching) or unobserved ones (e.g., instrumental variables) (Stukel, Fisher, Wenenberg, *et al.*, 2007). For example, the Energy Cost Indicator study (Hutton, Mauser, Fillitrault, *et al*., 1986) divided participants into quartiles according to annual energy consumption, then randomly assigned 100 randomly sampled participants from each quartile to each group (see also

Puckett *et al.,* 2004).

Sequence concealment bias can be avoided by hiding the sequence from researchers and participants, perhaps using third party central randomization (Higgins, Altman and Sterne, 2011) or the sequentially numbered, opaque, sealed envelopes (SNOSE) sometimes used in by medical researchers.

Attrition bias can be reduced with extrinsic incentives (e.g., completion bonuses) or intrinsic ones (e.g., such as stressing the value of complete data sets). Researchers can also adjust their data for attribution with intention-to-treat analysis (Hollis and Campbell, 1999), which uses imputation methods to adjust the partial data from participants who drop out, so that they are not lost altogether (Ibrahim, Chen, Lipsitz, *et al.,* 2005; Seligman, Darley and Becker, 1978). See PG&E's Smart-Rate pilot (George *et al*., 2010) or Idaho Power's Energy Watch Pilot (Kline, 2007).

Blinding bias can be reduced by providing as little information as possible, consistent with practical constraints (e.g., providing instruction, ensuring informed consent). When it is impossible to blind participants to condition (e.g., if they received an IHD) it might be possible to blind them to the other conditions. For example, the Milton Hydro experiment (Schembri, 2008) sought to limit contact between participants in its different conditions. CL&P's Plan-it Wise (Faruqui and Sergici, 2009b) randomly assigned participants without mentioning alternative conditions or allowing them to switch interventions if they did learn. See also the Ameren Illinois Power-Smart Pricing

(PSP) pilot (Violett *et al.*, 2010), BG&E Smart Energy Pricing Pilot (SEPP) (Faruqui and Sergici, 2009a), Pepco's PowerCents DC pilot (King, 2010), and Hydro Ottawa's Ontario Energy Boards Smart Price Pilot (Strapp, King and Talbott, 2007).

Volunteer selection bias can be avoided by requiring participation or by using an opt-out strategy, knowing that only strongly motivated individuals will change their default status. For example: (a) The Polk's Landing study (McClelland and Cook, 1979) installed in-home displays before residents purchased their homes. (b) The Southern California Edison experiment (Sexton, Johnson, and Konayama, 1987) had an opt-out design that was never used, both because the "exemption procedure was not well known" (p. 57) and because they were offered $100 to alleviate any financial hardship. Failing that, sample selection bias statistical methods (Heckman, 1976, 1979) may provide useful adjustments. For example, PG&E's Smart-Rate Pilot (George *et al.*, 2010) used propensity score matching to address differences between those who did and did not volunteer.

Our reported research makes three contributions. It assesses the degree of bias in field studies of electricity use. It provides meta-analyses of the results of existing studies, making alternative assumptions regarding their data. It provides guidelines for study design and directions for future methodological research. Future work should adapt reporting guidelines such as CONSORT to electricity field studies, design studies to avoid risk of bias, and conduct studies to quantify the impacts of bias in this domain. Its approach and

recommendations could be applied to field studies of any intervention, highlighting the need for general understanding of how these biases affect treatment effects.

4. Materials and Methods

We searched Google Scholar with the search terms: feedback + energy consumption, feedback + electricity + consumption, in-home feedback device + electricity, in-home display + pilots, real-time pricing, smart meter feedback devices, programmable thermostat, pricing program; in-home display; and automation. We also looked at the references of these studies and at publications citing them, as well as writing their authors asking about unpublished papers. Our search identified 112 potentially relevant papers of which 49 were eliminated for having no original data and another 31 for not satisfying our inclusion criteria: being in the US or Canada (11 studies), studying overall or peak reduction (11 studies), evaluating pricing, in-home displays or automatic controls (4 studies), and looking at residential use (2 studies). The supplementary materials (Appendix A) have details on the remaining 32 studies, 25 of which studied overall usage and 17 peak usage.

Two authors independently coded each study for risk of the six biases, using a method adapted from (Higgins, Altman and Sterne, 2011; Turner, Spigelhalter, Smith, *et al.*, 2009). Agreement on bias classification had high inter-rater reliability ($k = 0.75$) (Brennan and Prediger, 1981; Randolph, 2008). Our coding rules appear in Table 4.

[Insert Table 4]

Bias type	High risk	Low risk
Volunteer	• Opt-in design	• Opt-out design • Mandatory participation • Heckman Correction *
Intervention	• Random assignment before volunteering (allowing withdrawal) • Participant or researcher choice • Availability of intervention • Assignment based on pretests/baseline data	• Random assignment after volunteering • Propensity score *
Generation	• Alternating, day of birth, sequential, other non-random sequence	• Truly random sequence
Concealment	• Not low	• Central Randomization * • Sequentially numbered opaque sealed envelopes
Blinding	• Participants knew about other intervention groups when recruited.	• Participants were not informed about alternative intervention/control groups • Data collectors, analyzers, and writers blinded.
Attrition	• Data exclusions or withdrawals and data not missing at random	• Intention to treat analysis * • No dropouts or exclusions • Appropriate imputation methods *

As seen in Figure 2, the adequacy of the reporting varied considerably, as did the prevalence of bias -- where we could evaluate it: (a) All but one study reported whether participants had volunteered; all but four of those involved volunteers. (b) All studies reported whether the investigator or participant chose the intervention group; in roughly two-thirds of studies they did. (c) Only two

studies reported procedures for random assignment; one was proper and one not. (d) No study described its procedures for allocation concealment well enough to be evaluated; in correspondence, one author reported successful concealment. (e) Nine studies reported whether both participants and researchers were blinded to treatment group; in seven cases they were. (f) Half of studies reported attrition rates; most were represented high risk of bias.

[Insert Figure 2]

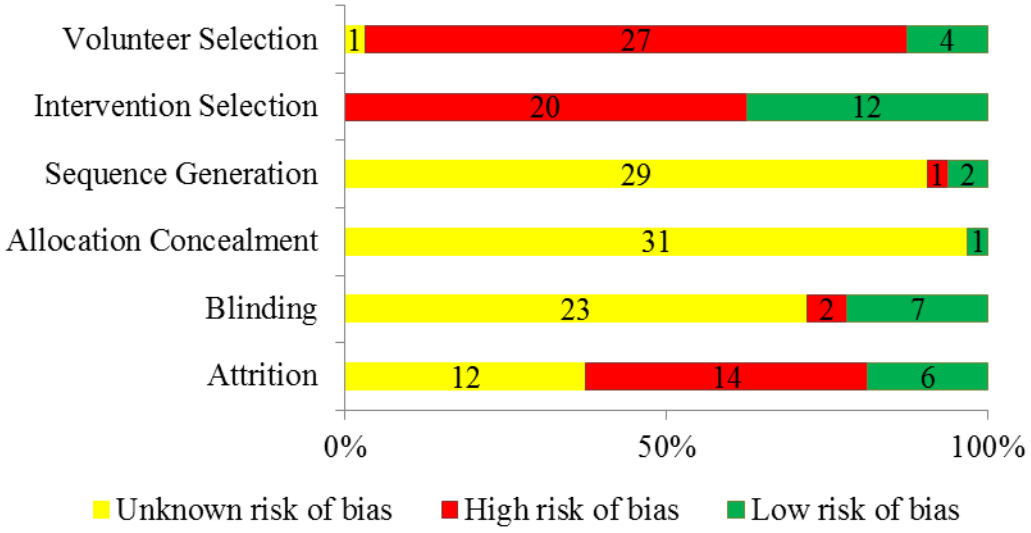

Table 1 shows Risk-of-Bias adjustment factors calculated from studies comparing medical clinical trials with and without biases (see Appendix B in supplementary materials). We omit quantification of volunteer bias since all studies have volunteer bias, and as a result nobody has bothered to quantify the magnitude of the bias. The adjusted treatment mean equals the observed mean divided by the mean adjustment factor. Thus, values greater than 1 indicate that the bias leads to overestimating mean effects. The adjusted variance is proportional to the sum of the unadjusted variance within

experimental conditions and the variance adjustment factor. Thus, variance adjustments greater than zero indicate that the bias leads to underestimating variance. When a study had more than one bias (which was always the case), the adjustment factors were multiplied, assuming that they were independent.

Acknowledgements

Financial support was received from the U.S. Department of Energy's Smart Grid Investment Grant (SGIG) funds and the Carnegie Electricity Industry Center. We thank Jack Wang for his valuable assistance, as well as Severin Borenstein, Ahmad Faruqui, Susan Frank, Stephen George, Laverne Gosling, Don Hammerstrom, Karen Herter, Kathryn Janda, John Kagel, Lou McClelland, Danny Parker, Mark Rebman, Clive Seligman, Richard Sexton, Brian Sipe, Dan Violette, Frank Wolak, and Tae-Jung Yun for providing in-depth information about their studies. The views expressed are those of the authors.

References

1. Abrahamse W, Steg L, Vlek C, Rothengatter T (2005) A review of intervention studies aimed at household energy conservation. *Journal of Environmental Psychology*, *25*(3): 273–291.
2. Ades A, Sutton A (2006) Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(1): 5–35.
3. Altman DG, Bland JM (1999) Treatment allocation in controlled trials: why randomise? *BMJ*, *318*(7192): 1209.
4. Barclay S, Todd C, Finlay I, Grande G, Wyatt P (2002) Not another questionnaire! Maximizing the response rate, predicting non-response and assessing non-response bias in postal questionnaire studies of GPs. *Family Practice*, *19*(1): 105-111.
5. Baladi MS, Herriges JA, Sweeney TJ (1998) Residential response to voluntary time-of-use electricity rates. *Resources and Energy Economics*, 20(3): 225–244.
6. Brennan RL, Prediger DJ (1981) Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement* 41: 687-699.
7. Callahan CA, Hojat M, Gonnella JS (2007) Volunteer bias in medical education research: An empirical study of over three decades of longitudinal data. *Medical education*, *41*(8): 746–753.
8. Carleton RA, Sanders CA, Burack WR (1960) Heparin administration after acute myocardial infarction. *The New England Journal of Medicine*, *263*: 1002-1005.
9. Carrol E, Hatton E, Brown M. (2009) Residential Energy Use Behavior Change Pilot. *Saint Paul, MN.: Office of Energy Security, Minnesota Department of Commerce*.
10. Darby S. (2001) Making it obvious: designing feedback into energy consumption. *Proceedings, 2nd International Conference on Energy Efficiency in Household Appliances and Lighting. Italian Association of Energy Economists/EC-SAVE programme*.
11. Darby S. (2006) The effectiveness of feedback on energy consumption: A review for DEFRA

of the literature on metering, billing and direct displays. *Environmental Change Institute, University of Oxford*, 1–21.

12. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C *et al*. (2003) Evaluating non-randomised intervention studies. *Health Technology Assessment*, 7(27): 1-186.
13. Ehrhardt-Martinez K, Donnelly KA, Laitner S. (2010). ACEEE, Advanced Metering Initiatives and Residential Feedback Programs: A Meta-Review for Household Electricity-Saving Opportunities. Report E105.
14. Faruqui, Sergici (2009a) BG&E Smart Energy Pricing Pilot (SEPP).
15. Faruqui, Sergici (2009b) Connecticut Light and Power (CL&P) impact evaluation of CL&P's Plan-it-Wise energy program.
16. Faruqui A, Sergici S, Sharif A (2010) The impact of informational feedback on energy consumption–A survey of the experimental evidence. *Energy*, *35*(4): 1598–1608.
17. Faruqui A, Hledik R, Sergici S (2009) Piloting the smart grid. *The Electricity Journal*, *22*(7): 55–69.
18. FERC (2009) A National Assessment of Demand Response Potential. *prepared by The Brattle Group, Freeman, Sullivan & Co., and Global Energy Partners, June*.
19. Fischer C (2008) Feedback on household electricity consumption: a tool for saving energy? *Energy Efficiency*, 1(1): 79–104.
20. Greenland S (2005) Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(2): 267–306.
21. Ioannidis J (2011) Commentary: Adjusting for bias: a user's guide to performing plastic surgery on meta-analyses of observational studies. *International Journal of Epidemiology*, 40: 777-779.
22. Gelman A, Hill J (2007) *Data analysis using regression and multilevel/hierarchical models* (Vol. 3). Cambridge University Press New York.
23. George et al (2010) PG&E Load Impact Evaluation of Pacific Gas and Electric Company's Time-Based Pricing Tariffs.

24. Grady D, Herrington D, Bittner V, Blumenthal R, Davidson M, Hlatky M, Hsia J, et al (2002) Cardiovascular disease outcomes during 6.8 years of hormone therapy. *JAMA: The Journal of the American Medical Association*, *288*(1): 49-57.

25. Hammerstrom DJ, Ambrosio R, Brous J, Carlon TA, Chassin DP, DeSteese JG, Guttromson RT, et al (2007) Pacific Northwest GridWise Testbed Demonstration Projects. *Part I. Olympic Peninsula Project*.

26. Hartling L, Bond K, Vandermeer B, Seida J, Dryden DM, et al. (2011) Applying the Risk of Bias Tool in a Systematic Review of Combination Long-Acting Beta-Agonists and Inhaled Corticosteroids for Persistent Asthma. PLoS ONE 6(2): e17242. doi:10.1371/journal.pone.0017242

27. Heckman JJ (1976) *The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models*. NBER.

28. Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, 153–161.

29. Higgins J, Thompson SG (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11): 1539–1558.

30. Higgins JPT, Altman DG, Sterne JAC (2011) Chapter 8: Assessing risk of bias in included studies in *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 (updated March 2011), eds Higgins JPT, Green S. Available from www.cochrane-handbook.org.

31. Hodis HN, Mack WJ, Lobo RA, Shoupe D, Sevanian A, Mahrer PR, Selzer RH, et al. (2001) Estrogen in the prevention of atherosclerosis. *Annals of Internal Medicine*, *135*(11): 939-953.

32. Hollis S, Campbell F (1999) What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ: The British Medical Journal*, *319*(7211): 670-674.

33. Hutton RB, Mauser GA, Filiatrault P, Ahtola OT (1986) Effects of cost-related feedback on

consumer knowledge and consumption behavior: A field experimental approach. *The Journal of Consumer Research*, *13*(3): 327–336.

34. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH (2005) Missing-data methods for generalized linear models. *Journal of the American Statistical Association*, *100*(469): 332–346.

35. IEA (2007). Time of Use Pricing and Energy Use for Demand Management Delivery.

36. Jüni P, Altman DG, Egger M (2001) Assessing the quality of controlled clinical trials. *BMJ: British Medical Journal*, *323*(7303): 42-46.

37. Karlowski TR, Chalmers TC, Frenkel LD, Kapikian AZ, Lewis TL, Lynch JM (1975) Ascorbic acid for the common cold. *JAMA: The Journal of the American Medical Association*, *231*(10): 1038-1042.

38. King C (2010) PowerCentsDC™ Program Effect of Pricing and Advanced Feedback. *Energy efficiency in domestic appliances and lighting*, 46.

39. Kline (2007) Idaho Power 2006 Analysis of the Residential Time-of-Day and Energy Watch Pilot Programs: Final Report. December, 2006.

40. McClelland L, Cook SW (1979) Energy conservation effects of continuous in-home feedback in all-electric homes. *Journal of Environmental Systems*, *9*(2): 169–173.

41. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, et al. (2010) CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology*, 63(8): e1–e37.

42. Moher D, Pham B, Jones A, Cook DJ, Jadad AR et al. (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *The Lancet*, *352*(9128): 609–613.

43. Myers RH, Montgomery DC, Anderson-Cook CM (2009) *Response surface methodology: process and product optimization using designed experiments*. John Wiley & Sons Inc.

44. Neenan, (2009) *Residential electricity use feedback: A research synthesis and economic framework*. Electric Power Research Institute.

45. Orne, M. T. (1962). On the social psychology of the psychological experiment: With

particular reference to demand characteristics and their implications. *American Psychologist*, *17*(11): 776-783.

46. Parsons HM (1974) What happened at Hawthorne? *Science*, *183*(4128): 922.
47. Petitti D (2004) Commentary: hormone replacement therapy and coronary heart disease: four lessons. *International Journal of Epidemiology*, *33*(3): 461-463.
48. Puckett *et al.* (2004) AmerenUE Residential TOU Pilot Study Load Research Analysis: First Look Results. February, 2004. RLW Analytics.
49. Randolph JJ (2008) *Online Kappa Calculator.* Retrieved February 5, 2012 , from http://justus.randolph.name/kappa.
50. Review Manager (RevMan) [Computer program]. Version 5.1. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2011.
51. Roberts S., Baker W. (2003) Towards effective energy information: improving consumer feedback on energy consumption, report, Centre for Sustainable Energy, Bristol.
52. Rosenthal R, Rosnow RL (1975) The volunteer subject. New York: Wiley.
53. Schembri J (2008) The Influence of Home Energy Management Systems on the Behaviours of Residential Electricity Consumers: An Ontario, Canada Case Study.
54. Schulz KF (1995a) Subverting randomization in controlled trials. *JAMA: the Journal of the American Medical Association*, *274*(18): 1456.
55. Schulz, KF (1995b) Unbiased research and the human spirit: the challenges of randomized controlled trials. *CMAJ: Canadian Medical Association Journal*, *153*(6): 783.
56. Schulz KF, Grimes DA (2002) Allocation concealment in randomised trials: defending against deciphering. *The Lancet*, *359*(9306): 614–618.
57. Shun-Shin M, Thompson M, Heneghan C, Perera R, Harnden A, *et al*. (2009) Neuraminidase inhibitors for treatment and prophylaxis of influenza in children: systematic review and meta-analysis of randomised controlled trials. *BMJ: British Medical Journal*, *339-348*.
58. Spees K, Lave LB (2007) Demand response and electricity market efficiency. *The Electricity*

Journal 20(3): 69–85.

59. Spiegelhalter DJ, Best NG (2003) Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modeling. *Statistics in Medicine*, *22*(23): 3687–3709.

60. Seligman C, Darley JM, Becker LJ (1978) Behavioral approaches to residential energy conservation. *Energy and Buildings*, 1(3): 325–337.

61. Sexton RJ, Johnson NB, Konakayama A (1987) Consumer response to continuous-display electricity-use monitors in a time-of-use pricing experiment. *The Journal of Consumer Research*, *14*(1): 55–62.

62. Strapp, King, Talbott (2007) Ontario Energy Board Smart Price Pilot Final Report. July, 2007.

63. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, *et al*. (2007) Analysis of observational studies in the presence of treatment selection bias: Effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *JAMA: the Journal of the American Medical Association*, *297*(3): 278-285.

64. Sulyma I, Tiedemann K, Pedersen M, Rebman M, Yu M (2008) Experimental Evidence: A Residential Time of Use Pilot. *Notes*, 100(6.33): 6–33.

65. Train, McFadden and Goett (1987) Consumer attitudes and voluntary rate schedules for public utilities. *Review of Economic Statistics* (69): 383-391.

66. Tune G (1964) Response preferences: A review of some relevant literature. *Psychological Bulletin*, *61*(4): 286-302.

67. Turner RM, Spiegelhalter DJ, Smith G, Thompson SG (2009) Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1): 21–47.

68. Tversky A, Kahneman D (1971) Belief in the law of small numbers. *Psychological Bulletin*, *76*(2): 105-110.

69. Violette et al (2010) Ameren Power Smart Pricing, 2009 Annual Report.

70. Weisser D. (2007) A guide to life-cycle greenhouse gas (GHG) emissions from electric supply technologies. *Energy* 32(9): 1543–1559.

71. Wolpert RL, Mengersen KL (2004) Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: effects of environmental tobacco smoke. *Statistical Science*, *19*(3): 450–471.

72. Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. The MIT press.

Figure Legends

Figure 1. A forest plot of unadjusted and adjusted Generic Inverse Variance meta-analysis for peak reduction for DP and automation, DP only, and in-home display and DP. The x-axis shows the % reduction and 95% confidence interval.

Figure 2. Distribution of studies that meet the criteria for high, low, or unknown risk of bias updated to reflect author responses.

Table Legends

Table 1. Risk of Bias Adjustments for Overall and Peak Reduction. Overall count is the number of interventions with that bias combination. Peak count is the number of interventions with that bias combination.

Table 2. Generic inverse variance and Hierarchical Linear Model estimates of adjusted and unadjusted effects from the five intervention combinations on peak reduction. For the GIV estimate, "missing" means that the within-group variances were not reported. For the HLM estimate, "missing" means that the intervention effect was not reported.

Table 3. Generic inverse variance and Hierarchical Linear Model estimates of adjusted and unadjusted effects from the five intervention combinations on

overall reduction. For the GIV estimate, "missing" means that the within-group variances were not reported. For the HLM estimate, "missing" means that the intervention effect was not reported.

Table 4. Criteria for classifying studies as high or low risk of bias. Notes (*): Heckman Correction (Heckman, 1976, 1979) statistically controls for factors affecting individuals' chance of being in the sample. Propensity scores statistically models factors that lead participants to choose an intervention program (Wooldridge, 2002; Gelman and Hill, 2007). Central randomization is done by a third party (Higgins, Altman and Sterne, 2011). Intention-to-treat analysis treats participant in terms of their originally treatment assignment, regardless of any subsequent exclusion, non-adherence, or withdrawal (Hollis and Campbell, 1999). Imputation estimates the values of missing data (e.g., by the mean from non-missing data (Ibrahim, Chen, Lipsitz, et al., 2005)).