Energy Policy 62 (2013) 401–409

Contents lists available at ScienceDirect

Energy Policy

journal homepage: www.elsevier.com/locate/enpol

Setting a standard for electricity pilot studies

Alexander L. Davis [a,*,1], Tamar Krishnamurti [a], Baruch Fischhoff [a,b], Wandi Bruine de Bruin [a,c]

[a] Department of Engineering and Public Policy, Carnegie Mellon University, USA
[b] Department of Social and Decision Sciences, Carnegie Mellon University, USA
[c] Centre for Decision Research, Leeds University Business School, UK

HIGHLIGHTS

- We conduct a meta-analysis of field studies of in-home displays, dynamic pricing, and automation on overall and peak use.
- Studies were assessed and adjusted for risk-of-bias from inadequate experimental design.
- Most studies were at high risk-of-bias from multiple sources.
- In-home displays provided the best overall reduction in energy use, approximately 3% after adjustment for risk-of-bias.
- Even after adjustment, automation approximately doubled the effectiveness of dynamic pricing on peak reduction from 6% to 14%.

ARTICLE INFO

Article history:
Received 7 June 2012
Accepted 27 July 2013
Available online 30 August 2013

Keywords:
In-home display
Dynamic pricing
Smart thermostat

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 of the impacts of these interventions, conducted in the US or Canada. We find that methodological problems are common in the design of these studies, leading to artificially inflated results relative to what one would expect if the interventions were implemented in the general population. Particular problems include having volunteer participants who may have been especially motivated to reduce their electricity use, letting participants choose their preferred intervention, and having high attrition rates. Using estimates of bias from medical clinical trials as a guide, we recalculate impact estimates to adjust for bias, resulting in values that are often less than half of those reported in the reviewed studies. We estimate 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 (~11% reported; ~6% adjusted), especially when used in conjunction with home automation (~25% reported; ~14% adjusted). We conclude with recommendations that can improve pilot studies and the soundness of decisions based on their results.

1. Introduction

Reducing overall residential electricity consumption means lowering emissions of greenhouse gases and other pollutants (Weisser, 2007) and decreasing the need for additional power plants and transmission capacity (Faruqui et al., 2009). Reducing residential electricity consumption during peak demand times (e.g., hot summer afternoons) means lowering 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 consumption by changing consumer behavior.

Three common interventions for reducing overall and peak electricity use are (a) in-home displays that provide feedback about electricity consumption and prices; (b) dynamic pricing programs, where residential electricity prices more closely follow the wholesale market, creating an incentive to reduce use during peak-demand hours; and (c) automation, with programmable thermostats, smart switches, and similar technologies that control electricity use according to user specifications and electricity prices.

Although many studies have evaluated the effectiveness of such interventions, their designs and reporting protocols vary so much that it is hard to aggregate their results. Here, we propose a

[*] Corresponding author. Tel.: +1 412 216 2040.
E-mail address: alexander.l.davis1@gmail.com (A.L. Davis).
[1] All data and materials can be found at Harvard's Dataverse; http://hdl.handle.net/1902.1/21586

0301-4215/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.enpol.2013.07.093

standard approach, based on the *risk-of-bias* methodology that has been developed to analyze results from medical clinical trials (Higgins et al., 2011; Moher et al., 2010), and has been applied to treatments as diverse as inhaled corticosteroids for asthma (Hartling et al., 2011), routine antenatal care (Turner et al., 2009), and influenza treatment and prevention (Shun-Shin et al., 2009). We draw on this methodology to analyze electricity field trials in the US and Canada for 6 common types of bias, adjust reported effects for bias, and aggregate those adjustments to produce revised estimates of study effects.

Risk-of-bias analysis arose from medical researchers' observation that methodologically flawed studies often found positive health effects that vanished, or were reversed, in sounder studies (Moher et al., 1998). For example, the initial promise of hormone replacement therapy (Petitti, 2004) was later found to reflect a selection bias, whereby women who opted for the therapy had relatively high socio-economic status, which is associated with better health outcomes (Grady et al., 2002). In order to account for such problems, risk-of-bias analysis first characterizes studies in terms of known biases in their design and execution, and then adjusts reported effect sizes for the expected impact of each bias on study results. The adjustments are based on studies of the differences observed in clinical trials conducted with and without a bias.

We follow the same logic in analyzing the effects of interventions on residential electricity use. We use correction factors from medical research as a starting point, in the absence of ones for electricity field trials. Our concluding discussion considers why the corresponding biases might be larger or smaller in electricity field trials, as well as how electricity researchers might create analogous estimates, as the discipline builds up its methodological and empirical base.

Risk-of-bias analysis is only possible when a study is reported with sufficient detail to evaluate its procedures. Reviews have, however, often noted problems with the reporting of electricity field studies (Fischer, 2008). Indeed, Abrahamse et al. (2005) concluded that reporting was so deficient that a “thorough meta-analysis was not deemed feasible” (p. 275). Our review encountered similar problems. As a result, we also propose ways to deal with deficient reports and improve future reporting practices. We believe that following these procedures will allow utilities, regulators, investors, and others to make sounder decisions based on trial results.

1.1. Six common biases

The Cochrane Collaboration[2] conducts continuing systematic reviews of medical treatments (Higgins et al., 2011). Each review requires risk-of-bias assessments for all included studies. Based on the risk-of-bias categories used by the Cochrane Collaboration, we identified six common biases that are likely to have serious effects in electricity research: (1) *volunteer selection bias*, arising when individuals who chose to participate in a study might be more likely to change their behavior than individuals drawn from the target population in a representative way; (2) *intervention selection bias*, arising when participants decide how to participate, by selecting their preferred experimental condition; (3) *sequence generation bias*, arising when a non-random procedure assigns participants to conditions; (4) *allocation concealment bias*, arising when experimenters or participants know the assignment sequence, hence might manipulate the assignment to condition; (5) *blinding bias*, arising when researchers' knowledge of group assignment affects their treatment of participants or their interpretation of participants' behavior; and (6) *attrition bias*, arising when participants who do not benefit from the study are excluded or withdraw from the study before its conclusion. Each bias has potential analogs in electricity field trials, threatening the validity of their results.

Volunteer selection bias arises when the people who choose to participate in a study differ from the study's intended population. Research in other domains has found that volunteers differ from randomly sampled individuals in ways that might bias treatment effects (Barclay et al., 2002; Callahan et al., 2007; Rosenthal and Rosnow, 1975). An example in energy research is Sulyma et al.'s (2008) finding of relatively high education and income levels among British Columbia Hydro customers who volunteered for pilot tests of dynamic pricing and in-home displays. If such individuals are also better able to comprehend and respond to program information, compared to average consumers, then studies involving them will overestimate program impacts. Baladi et al. (1998) found evidence of just such a bias: volunteers were better than non-volunteers at estimating peak-demand electricity use in a time-of-use experiment and more optimistic about program benefits. Of course, volunteer selection bias might not be a problem if the actual program recruited participants in the same way as the trial (e.g., by enrolling only relatively wealthy, well-educated, and optimistic consumers). Researchers who recognize this threat are sometimes tempted to compare volunteers and non-volunteers on observable characteristics (e.g., demographics, baseline electricity usage), interpreting a lack of difference on these variables as indicating no volunteer bias. However, individuals matched in these ways need not respond similarly in the study, if they differ in other relevant ways that are not represented in the available measures or if the act of volunteering changes their behavior (Davis and Krishnamurti, 2013).

Intervention selection bias occurs when, once in a study, participants choose their treatment group, rather than being randomly assigned (Altman and Bland, 1999b). In the electricity context, this bias would lead to overestimating an intervention's effectiveness, if people who select it are especially motivated to change their behavior or to make the intervention work. Hammerstrom et al. (2007) assigned participants randomly to control or tariff groups, but then allowed those receiving tariffs to choose among three options: fixed rate, time of use with critical peak pricing, and real-time pricing. If participants chose the plan that best suited them, then the study's results would overestimate the effects of universal adoption of any of the three options. For example, if people who can most easily shift their electricity use to off-peak hours chose real-time pricing, then one would expect smaller reductions (and more objections) were all consumers enrolled in that plan, including consumers who cannot make that shift (e.g., because their health requires air conditioning during peak usage times). Here, too, the bias is not a threat to generalization if, once enrolled in the program, all consumers will choose options in the same way as those in the trial.

Sequence generation bias occurs when a non-random process assigns participants to conditions (e.g., alphabetically or by alternation) (Altman and Bland, 1999a; Schulz and Grimes, 2002). For example, Faruqui and Sergici (2009b) first recruited consumers for dynamic pricing, then for peak-time rebate, and so on through their study's 24 conditions. If people who are most eager to receive some intervention sign up first, then serial enrollment may overestimate the effectiveness of the interventions assigned first. Randomization must use formal procedures because people have difficulty generating random sequences on their own (Tune, 1964; Tversky and Kahneman, 1971) and may not even realize what it entails. In one review, Schulz and Grimes (2002) found that 5% of medical clinical trials reporting random assignment actually used

[2] http://www.cochrane.org/

deterministic rules (e.g., alternation, date of birth, day of hospital admission). An additional 63% reported too few details to determine their procedure.

Allocation concealment bias occurs when participants or researchers know the assignment sequence (even if random) and can use that knowledge to manipulate individual participants' assignment (e.g., getting favored patients into non-placebo groups; postponing medical treatment until a clinical trial begins). Schulz (1995a, 1995b) provides anecdotes of researchers who tried to decipher allocation sequences, for example, by holding sealed assignment envelopes up to a light to see their contents (Carleton et al., 1960). In an electricity field study, such a bias might involve favoring larger houses when assigning a smart thermostat, as a result of thinking that those households will derive the most benefit or responding to pressure from their owners for special treatment.

Blinding bias occurs when the treatment assignment is known by researchers, who might, for example, give better care to patients in the treatment group, or by participants, who might, for example, feel neglected in a control group (Miller and Stewart, 2011). Hutton et al. (1986) found that consumers told that they were in an energy consumption study reduced their electricity use more than those who were not told (−310 kWh vs. −270 kWh per month, respectively), even though neither group received an actual intervention. That difference might reflect a Hawthorne effect, where just knowing that one is being studied changes behavior (Orne, 1962; Parsons, 1974). In a study on the effects of ascorbic acid on the common cold, Karlowski et al. (1975) found that some participants guessed their condition based on the taste of their pills. Those in the placebo group who guessed correctly reported more severe symptoms and were more likely to drop out of the study, compared to those who thought that they had received ascorbic acid. Correcting for these problems eliminated what little evidence there was for protective effects of ascorbic acid.

Attrition bias occurs when participants withdraw or are excluded from a study for reasons related to their assigned intervention (Hollis and Campbell, 1999). The inadequate blinding in the ascorbic acid study led to such an attrition bias, in addition to affecting symptom reports. In an example from electricity research, a pilot study looking at different combinations of dynamic pricing, in-home displays, and automation reported a 1% monthly dropout rate for unknown reasons (Faruqui and Sergici, 2009a). An intervention's effect will be overestimated if people who are getting no benefit are more likely to withdraw, thereby leaving no record of its ineffectiveness for them. The same (or the opposite) could be true for people who drop out because they move, die, have problems that hamper their participation, or fall out with researchers. Even when conditions have equal attrition rates, the causes of attrition may be different, leading to biased results.

2. Materials and methods

2.1. Selecting studies for analysis

To find relevant studies, we searched Google Scholar with the following 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. For each study that the search retrieved, we examined its references and the publications citing it, as well as wrote the authors to inquire about unreported details. 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 (eliminating 11 studies); studying overall or peak reduction (11); evaluating some combination of pricing, in-home displays or automatic controls (4); and looking at residential use (2). Appendix A has details on the remaining 32 studies, 26 of which studied overall usage and 18 peak usage.

2.2. Coding for bias

Two authors independently coded each study for risk of each of the six biases, using a method adapted from previous research (Higgins et al., 2011; Turner et al., 2009) and with the coding rules summarized in Table 1. The two coders had high inter-rater reliability ($\kappa = .75$) (Brennan and Prediger, 1981).

2.3. Estimating effects in field trials

The primary outcome of interest was percent reduction in electricity consumption for residential electricity customers, as observed overall or during peak-demand hours. When studies did

Table 1
Criteria for classifying studies as high or low risk-of-bias.

Bias type	High risk	Low risk
Volunteer	Opt-in design	(1) Opt-out design (2) Mandatory participation (3) Heckman correction[a]
Intervention	(1) Random assignment before volunteering (allowing withdrawal) (2) Participant or researcher choice (3) Availability of intervention (4) Assignment based on pretests or baseline data	(1) Random assignment after volunteering (2) Propensity score adjustment[b]
Generation	Alternating, day of birth, sequential, other non-random sequence	Truly random sequence
Concealment	Not central randomization or similar procedure	Central randomization[c]
Blinding	Participants knew about other intervention groups when recruited	Participants were not informed about alternative intervention or control groups
Attrition	Data exclusions or withdrawals, and data not missing at random	(1) No dropouts or exclusions (2) Intention-to-treat analysis[d] (3) Appropriate imputation[e]

[a] The Heckman (1979) correction statistically controls for factors affecting individuals' chance of being in the sample.

[b] Propensity score adjustment statistically models factors that lead participants to choose an intervention program (Gelman and Hill, 2007; Wooldridge, 2002).

[c] Central randomization is done by a third party (Higgins et al., 2011).

[d] Intention-to-treat analysis treats participant in terms of their original treatment assignment, regardless of any subsequent exclusion, non-adherence, or withdrawal (Hollis and Campbell, 1999).

[e] Imputation estimates the values of missing data (e.g., by the mean from non-missing data, Ibrahim et al., 2005).

not present this number directly, we calculated it from average kWh use, using the control group as a base. For the sake of comparability across studies, we focused on the most commonly reported measures, pooled participants across conditions and study phases (which were often divided differently across studies), and used annual averages. We used the squared standard error of the reported effect in each group as an estimate of within-group variation. When standard errors were not reported, we attempted to determine them from reported *t* values or standard deviations and sample sizes. (Section 1.1 of the Supplementary Information has further details on data treatment.)

2.4. Correcting for high risk of individual biases

Researchers who conduct systematic reviews have examined the degree to which high or unknown risk-of-bias is associated with over- or underestimation of reported results. For example, Jüni et al. (2001) found that studies with inadequate or unclear allocation concealment reported 30% larger intervention effects than did studies with adequate concealment. The 95% confidence interval (CI) around the bias estimate was between 20% and 38% (denoted as 95% CI: [20%, 38%]). The confidence interval can then be used to approximate the variance of the bias estimate, which, in the case of allocation concealment, is 25% ($=((30-20)/2)^2$).

Table 2 shows our mean estimates for five of the six biases (in % increase or decrease relative to an unbiased study), bracketed by 95% confidence intervals, drawn primarily from four major meta-epidemiological studies (Jüni et al., 2001), supplemented by additional analyses of our own. There is no estimate for volunteer selection bias, given that there is no general way of knowing how the behavior of people who volunteer for a study differs from that of people who do not.

Our estimate of intervention selection bias is based on Stukel et al. (2007), who analyzed observational data comparing patients who received cardiac catheterization for acute myocardial infarction compared to those who did not. The study reported a risk ratio of 0.37 (95% CI: [0.35, 0.38]) before adjusting for intervention selection bias, suggesting a very large improvement (where a risk ratio of 1.0 reflects no change, and lower risk ratios indicate larger improvements). However, that improvement shrank considerably when the results were adjusted for selection bias using two different statistical methods. One method used an instrumental variable rather than the actual treatment to estimate the effect of cardiac catheterization (Wooldridge, 2002); where an instrumental variable is highly correlated with the treatment variable (here, cardiac catheterization), but uncorrelated with factors that cause the outcome (heart attacks), such as a history of congestive heart failure. The second method uses a propensity score (Gelman and Hill, 2007) as a covariate; this score estimates each person's probability of choosing the treatment group over the control, based on other measured variables.

Table 2
Estimates of bias for each of the six types of bias.

Source	Bias type	Bias estimate[a] (%)	95% CI	Variance[b] (%)
–	Volunteer	–	–	–
Stukel et al. (2007)	Intervention	44	[19, 69]	156
Jüni et al. (2001)	Generation	19	[−9, 40]	196
Jüni et al. (2001)	Concealment	30	[20, 38]	25
Jüni et al. (2001)	Blinding	14	[1, 26]	43
Kjaergard et al. (2001) Schulz et al. (1995)	Attrition	−8	[−21, 6]	49

[a] *Bias estimate* is the expected value of the bias estimate.

[b] *Variance* is the variance of the bias estimate based on the 95% confidence intervals.

Using the instrumental variable method, Stukel et al. (2007) derived an unbiased risk ratio of 0.86 (95% CI: [0.78, 0.94]), indicating a much smaller treatment effect than the 0.37 risk ratio reported in the study. We calculated the ratio of the reported risk ratio to the instrumental variable- adjusted risk ratio of 0.43 (=0.37/0.86) (95% CI: [0.35, 0.51]), or 57% overestimation (=(0.37–0.86)/0.86). Using propensity-score adjustment produced a risk ratio of 0.54 (95% CI: [0.53, 0.55]), with the corresponding ratio of reported risk ratio to propensity-score adjusted risk ratio of 0.69 (=0.37/0.54) (95% CI: [0.65, 0.73]), or 31% overestimation (=(0.37–0.54)/0.54). Combining the two adjustments with a random-effects meta-analysis produces an aggregate bias estimate of 0.56 (95% CI: [0.31, 0.81]), equal to 44% overestimation (=(0.56−1)/1). Our adjustment for intervention selection used that value. For example, if a study reported 40% peak reduction but was at high risk for intervention selection bias, the adjusted peak reduction was 28% (=40%/1.44).

Our estimates of the magnitude of bias for sequence generation, allocation concealment, and blinding bias were taken from Jüni et al. (2001), who aggregated four systematic reviews of clinical trials that compared the odds ratios of trials with high (or unclear) risk of bias against the odds ratios of studies with low risk of bias. They found (a) that studies with inadequate or unclear sequence generation had a 19% larger estimate of intervention effects (based on odds ratios) than did studies with adequate sequence generation (95% CI: [−9%, 40%]), (b) that studies with inadequate or unclear allocation concealment had a 30% larger estimate of intervention effects than did those with adequate allocation concealment (95% CI: [20%, 38%]), and (c) that studies with inadequate or unclear blinding had a 14% larger estimate of intervention effects than did those with adequate blinding (95% CI: [1%, 26%]). Thus, the bias adjustment factors are 1.19, 1.30, and 1.14, respectively.

Our estimate of the magnitude of attrition bias was taken from two reviews. In the first, Kjaergard et al. (2001) found an underestimation of 50% (95%CI: [−178%, 20%]) for studies with improper follow-up (which includes attrition both while a treatment is being applied and in subsequent measurement of its effects). In the second, Schulz et al. (1995) found 7% underestimation of effects for studies that excluded eligible participants from statistical analyses due to protocol deviations, withdrawals, drop-outs, and losses to follow-up, compared to studies that included all participants (95%CI: [−21%, 6%]). Combining these two reviews with a random-effects meta-analysis produces an aggregate bias estimate of −8% (95%CI: [−21%, 6%]). This gives a bias adjustment factor of 0.92.

The medical research literature conventionally reports results in terms of risk (or odds) ratios, comparing results in treatment and control conditions of categorical outcomes (e.g., cured/not cured, improved/not improved, alive/dead) (Bland and Altman, 2000). Thus, an odds ratio of 1.3 might mean that 30% more patients were cured in the treatment group than in the control group. If a bias leads to overestimating that treatment's odds ratio by 50%, then an odds ratio of 1.15 would be a better estimate of the treatment's effect (using the calculation described above). Odds and risk ratios do not translate directly into the continuous effects measured in electricity trials (unless passing a threshold, such as a state-mandated reduction, is considered a successful treatment). We have adopted the convention of interpreting a 50% overestimation in odds or risk ratios as equivalent to a 50% overestimation of reductions in electricity use for studies with that bias.

When a study had more than one bias (as was always the case), we multiplied the adjustment factors for those biases, assuming that they were independent, again following Turner et al. (2009). For example, a study with intervention selection and sequence generation biases would have its reported effect (in percent reduction in kWh) divided by 1.71 (=1.44 × 1.19). The bias adjustment also changes the variance of the reported effect (see Eqs. (7)–(9) in Section 2 of the Supplementary Information).

2.5. Aggregating results across studies

We focus on two general approaches to aggregating intervention effects across studies. The first approach uses Generic Inverse Variance (GIV) meta-analysis, which weights each study's reported intervention effect by the inverse of its *within-group* variance, defined as the squared standard error of the reported effect for each group in the study. GIV gives more weight to groups with more stable estimates (i.e., with smaller within-group variance and larger sample sizes) (Thompson and Higgins, 2002). We use a random-effects model so as to accommodate *between-group* variation, treating some of the difference between groups that received the same treatment as reflecting true differences in effect size rather than just sampling variation (Hedges and Vevea, 1998). If between-group variation is large (because the same treatment produces very different results), the importance of within-group variance is reduced, so that all groups receive roughly equal weight. If between-group variation is small, then the random-effects model weights each study roughly by the inverse of its within-group variance.

The second approach uses Hierarchical Linear Models (HLM) (Gelman and Hill, 2007), which treat each group's results as a single observation, making it possible to include studies that did not report standard errors (as was often the case here). When a study has multiple groups receiving variants of the same treatment (e.g., different IHDs), HLM estimates a study-level intercept that accounts for the *within-study* variation, observed as correlations between estimates in the same study. (Additional detail about GIV and HLM approaches is provided in of the Supplementary Information.) We also used Ordinary Least Squares (OLS) aggregation, which also ignores within-group variance, but does not take within-study variation into account. However, as it revealed similar patterns and is inferior to HLM, we do not report those results.

3. Results

3.1. Prevalence of high or unknown risk of bias

Fig. 1 shows the results of coding each study in terms of whether it provided enough information to determine whether it was subject to each bias and, if so, whether it had a high or low risk of bias, using the coding scheme in Table 1. As seen in the figure, the adequacy of the reporting varied considerably, as did the prevalence of bias—in studies whose reporting allowed us to evaluate it: (a) all but one study reported whether participants had volunteered; all but four of those involved volunteers (coded as high risk of bias). (b) All studies reported whether the investigator or participant chose the intervention group; in roughly two-thirds of studies they did (coded as high risk of bias). (c) Only two studies reported procedures for random assignment; one was truly random, the other was 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 participants were blind to alternative treatment groups; in seven cases they were. (f) Half of the studies reported attrition rates; most had high risk of bias.

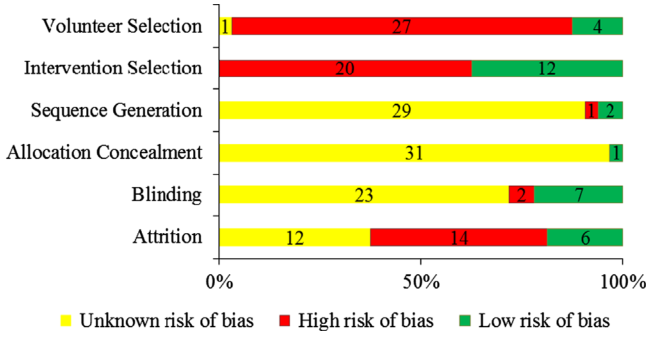

Fig. 1. Distribution of studies that meet the criteria for high, low, or unknown risk-of-bias updated to reflect author responses.

3.2. Correction for risk of biases

The first column of Table 3 shows the combinations of bias observed in these studies, grouping ones with high or unknown risk of bias, following Jüni et al. (2001) in assuming that there was high risk when studies failed to report procedures that controlled for a bias. All studies had high risk of at least two biases, sometimes more. The second column, labeled *Estimate*, shows the bias adjustment for each combination, computed as the product of the adjustment factors for the individual biases (Table 3). The third column, labeled *Variance*, shows the variance of the bias adjustments for the treatment groups involved. The fourth and fifth columns indicate the number of treatment groups with each combination of bias, for overall and peak usage. For example, the first row shows that the three treatment groups in which overall use was measured had high risk of both allocation concealment bias (4) and attrition bias (6). As a result, we adjusted the treatment effect for each group by dividing by 1.20 (=0.92 × 1.30) and increasing its variance by $\sigma_i^2 = 0.017$ (see Eq. (8) in Section 2 of the Supplementary Information). The same adjustments were made for the three treatment groups in which reductions in peak energy use were measured, all of which had high risk of these two biases.

Table 4 summarizes the effects for all interventions targeting residential electricity use during peak hours. With both aggregation methods (GIV and HLM), dynamic pricing alone (*Pricing Only*) produces significant reductions, which are roughly halved by the risk-of-bias adjustment. Combining dynamic pricing with automation (*Pricing and Auto*) substantially increases those effects. Adding in-home displays, however, contributes little to the effectiveness of dynamic pricing (*IHD and Pricing*), or when combined with automation (*All Three*).

Table 5 reports comparable statistics for reductions in overall residential electricity use. Aggregation with GIV found the greatest reductions with in-home displays alone (*IHD Only*), equal to about

Table 3
Combinations of risk-of-bias adjustments from the 32 studies.

Bias combinations[a]	Adjustment		# of groups	
	Estimate	Variance	Overall	Peak
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

[a] 1=volunteer bias, 2=intervention selection bias, 3=sequence generation bias, 4=allocation concealment bias, 5=blinding bias, 6=attrition bias. Although most studies had volunteer bias, it does not appear in column 1 of the table because estimates of bias magnitude could not be found.

Table 4
Generic inverse variance (GIV) and hierarchical linear model (HLM) estimates of reported and adjusted effects on peak reduction for four of the five intervention combinations.[a]

Intervention	Generic inverse variance			Hierarchical linear model		
	Reported	Adjusted	Groups[b]	Reported	Adjusted	Groups[c]
Pricing only	**6.93***	**2.49***	11	**10.83***	**5.92***	28
	(1.74)	(0.64)	(17)	(1.77)	(1.21)	(3)
IHD and pricing	6.30	3.06	1	**12.91***	**6.17***	10
	(4.34)	(4.44)	(10)	(4.44)	(2.11)	(1)
Pricing and automation	**32.80***	**16.81***	3	**25.35***	**13.83***	16
	(5.50)	(3.60)	(16)	(4.44)	(2.65)	(3)
All three	–	–	0	**31.30***	**12.56***	3
			(4)	(1.41)	(3.30)	(1)

[a] Standard errors are in parentheses.

[b] For the GIV estimate, the number in parentheses in the Groups column is the number of groups that did not report within-group variances or an intervention effect, hence could not be included.

[c] For the HLM estimate, the number in parentheses in the Groups column is the number of groups that did not report an intervention effect, making bias estimation and correction impossible.

* indicates $p < 0.05$.

Table 5
Generic inverse variance (GIV) and hierarchical linear model (HLM) estimates of reported and adjusted effects on overall reduction for the five intervention combinations.

Intervention	Generic inverse variance			Hierarchical linear model		
	Reported	Adjusted	Groups	Reported	Adjusted	Groups
IHD only	**3.76***	1.12	12	**5.10***	**2.99***	16
	(1.27)	(1.11)	(5)	(2.33)	(1.20)	(0)
Pricing only	**2.84***	0.30	10	1.73	0.91	21
	(0.92)	(0.35)	(18)	(1.17)	(0.61)	(10)
IHD and pricing	2.17	1.05	1	2.30	1.32	8
	(3.51)	(2.94)	(9)	(1.41)	(0.87)	(3)
Pricing and automation	**3.60***	0.09	4	3.84	2.81	11
	(1.58)	(0.14)	(14)	(2.21)	(1.92)	(7)
All three	**3.00***	1.28	1	3.00	1.28	1
	(0.323)	(2.18)	(3)	–	–	(3)

4% in the reported data and 1% after adjustment for risk-of-bias. HLM analysis also found in-home displays alone to be most effective, reducing usage by about 5% in the reported data and 3% after bias adjustment. Although all five interventions show lower overall usage after risk-of-bias adjustment, that difference is statistically significant only for in-home displays alone.

As mentioned, none of these adjustment include volunteer bias, which was present in most studies (Fig. 1), but for which no adjustment estimates are available.

4. Discussion

We report a critical meta-analysis of 32 publicly available evaluation studies estimating the effectiveness of three 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 developed for medical clinical trials, we find 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. Estimates of the magnitude of bias from risk-of-bias research were used to adjust the reported effects. We aggregated estimates from individual studies using two meta-analytic procedures, Generic Inverse Variance (GIV) and Hierarchical Linear Models (HLM), which produced similar patterns but somewhat different estimates of effect size.

For peak electricity use, risk-of-bias adjustment roughly halved the observed effect size (Table 4). Using both reported and adjusted estimates, both GIV and HLM found that dynamic pricing produced statistically significant savings. Those effects increased markedly when dynamic pricing was paired with automation, but not with IHDs—which also added nothing to the combination of dynamic pricing and automation together. Thus, the evidence suggests that automation responds effectively to dynamic pricing schemes, whereas asking consumers to monitor a display does not. Unfortunately, no study examined the effects on peak load reduction of automation or IHDs alone, without dynamic pricing.

For overall electricity use (Table 5), most interventions showed statistically significant reductions when applying the more sensitive GIV aggregation method to reported estimates, with the exception of dynamic pricing with IHDs. However, risk-of-bias adjustments reduced all reported effects to 3% or less, leaving but one statistically significant result: in-home displays alone, when aggregated with HLM (but not GIV). Dynamic pricing alone, which significantly reduced peak use, had weak effects on overall use. It seemed to reduce the usefulness of IHDs, a combination that was also no more effective than dynamic pricing alone with peak use, possibly reflecting cognitive overload for consumers.

Thus, without risk-of-bias adjustments that correct for potential methodological flaws, many interventions seem to reduce overall usage. However, with those adjustments, only IHDs did (by about 3%). For peak usage, dynamic pricing, especially when combined with automation, is effective either way, although risk-of-bias adjustments halve the effect size.

Almost none of the 32 studies reported enough information to assess their vulnerability to methodological problems due to sequence generation, allocation concealment, and blinding bias; about half lacked details needed to assess attrition bias. Although that lack of information might indicate that their use of appropriate methods went without saying, risk-of-bias analyses have typically found that studies with unknown risk-of-bias tend to report larger effects than do studies known to have low risk-of-bias. As a result, we treated studies with inadequate reports as having high risk of bias, which may have meant unwarranted adjustments in studies that had sound procedures, but failed to describe them. Conversely, we made no adjustment for volunteer bias in the 27 studies that reported relying on individuals who selected themselves into the test (and the one that said nothing).

Another common reporting problem emerged with the results. Although most studies reported overall effects, many did not report within-group variances. As a result, we could not apply the more sensitive GIV aggregation method to them.

An important limitation of our analyses is that our adjustment factors are taken from studies of medical clinical trials, which use different treatments, participant populations, research methods, dependent variables, and statistical estimates. We also needed to equate the categorical odds ratios of clinical trials to percentage changes in electricity consumption. Although that translation should still show the relative impact of different biases, a 50% ratio may not mean a 50% change. There is a vital need for studies creating comparable adjustment factors for electricity field trials. The same mechanisms seem plausible (e.g., without adequate controls, treatments will be allocated to people who seem most likely to benefit from them, thereby exaggerating their effect in the general population).

However, in the absence of evidence, their magnitude is a matter for speculation. For example, it may be easier to blind researchers and participants in medical clinical trials, where placebos are routinely used, than in electricity field trials, which often require a visible device or incentive program. As a result, blinding bias may be larger with field trials. Patients may be more motivated than electricity consumers to volunteer for trials, increasing volunteer bias in medicine. On the other hand, medical trials often have strong scientific, legal, and ethical pressures to ensure equal access, reducing volunteer bias. Unless electricity field trials can be conducted without biases, we need evidence about their impact, in order to interpret results appropriately.

A second limitation of our study is its assumption of independence when combining the bias adjustment factors. As seen in Table 3, studies weak in one way also tended to be weak in others. However, we have no direct evidence on how the resulting biases combine. As a result, we assumed that the effects compounded, rather than mitigating or exacerbating, one another. Here, too, evidence is needed.

Finally, our conclusions are limited to existing studies. Even with 32 studies, we were forced to combine interventions into classes whose members might have varied in their effectiveness (e.g., dynamic pricing programs with different incentives; feedback delivered with different frequencies and displays). We also lacked the combinations of interventions needed to clarify their joint and separate effects. For example, we cannot untangle how IHDs help or hinder the effect of dynamic pricing on peak use, or how well automation does by itself. Studies that vary consistently in ways beyond the interventions being tested are called *meta-confounded* (Deeks et al., 2003). Although we used two different procedures to aggregate results (GIV and HLM), and found generally similar results, other approaches are possible, especially if within-group variances were routinely reported (Ioannidis, 2011; Spiegelhalter and Best, 2003).

5. Recommendations

5.1. Reporting electricity field studies

The incomplete reporting of most field studies limited not only our analyses, but also their value to decision makers relying on them. A straightforward way to improve the reporting, and value, of field studies is to adapt the widely used CONSORT guidelines for medical clinical trials.[3] CONSORT provides a 25-item reporting checklist that includes the trial design (e.g., factorial); the participants and their eligibility criteria; the sequence generation, allocation concealment, and blinding processes; statistical methods, and open access to supplementary materials that allow anyone to replicate a study and its analyses. One item on the checklist is the within-group variances needed for the more powerful GIV aggregation.

Two reporting requirements specific to comparing and aggregating results from electricity field trials are: (a) effects on both overall and peak usage, in order to see how interventions with one goal affect the other, and (b) full usage statistics for all time periods for all intervention groups.

5.2. Designing electricity field studies

Risk-of-bias adjustments would not be needed, were studies conducted without bias. We recognize that practical concerns (e.g., Public Utility Commission approval) sometimes constrain a study's design. It is our perception that the quality of studies is improving over time, in part, through the growing number that appear in the peer-reviewed literature, which encourages stronger practices (and reporting). Guidelines proposed by researchers at the Lawrence Berkeley National Laboratory (Todd et al., 2012) and the Electric Power Research Institute (Neenan and Robinson, 2010) should further help to codify strong methodological practices. We conclude with recommendations for reducing each bias, including examples of successful designs.

Volunteer selection bias can be reduced either by requiring participation or by using an opt-out design (Junghans et al., 2005), assuming that only strongly motivated individuals will change their default status. For example, a study at Polk's Landing North Carolina installed in-home displays before residents purchased their homes (McClelland and Cook, 1979). A study of Southern California Edison consumers had an opt-out design, which lost no participant, partly because the “exemption procedure was not well known” (p. 57) and partly because participants were offered $100 to stay in the study (Sexton et al., 1987). Failing that, statistical methods can sometimes improve estimates (Heckman, 1979). For example, PG&E's Smart-Rate Pilot used propensity score matching to compare consumers who did and did not volunteer for the pilot (George et al., 2011).

Intervention selection bias can be reduced by assigning consumers to interventions randomly, after they have enrolled in a study (Altman and Bland, 1999b). For example, the Iowa Residential Electricity Study (RES) randomly assigned participants to time-of-use pricing or the control group after they had signed up to participate (Baladi et al., 1998). See also BC Hydro's AMI study (Sulyma et al., 2008), the Twin Rivers study (Seligman et al., 1978), and AmerenUE's Residential TOU pilot (Puckett and Hennessy, 2004).

Sequence generation bias is best reduced with formal randomization (Altman and Bland, 1999a). For example, Puckett and Hennessy (2004) used the “rand()” function in Microsoft Excel to generate a random number between 0 and 1, assigning participants with a number greater than 0.5 to the treatment group and those with a lower number to the control group.

Allocation concealment bias can be reduced by hiding the random sequence and group assignment from researchers and participants, perhaps by having a third party managing the process (Higgins et al., 2011). For example, a third party might create two types of “welcome package,” differing solely in whether they included a smart thermostat, thus concealing the allocation to both researcher and recipient.

Blinding bias can be reduced with *placebos*, mimicking the treatment in every way except the element being tested. For example, if a treatment group received an in-home display providing near real-time feedback on electricity consumption, a placebo control group might receive an identical display without the feedback. Whatever participants learn about their own condition, they should be told as little information as possible about the alternative conditions, lest they feel discouraged or resentful about not receiving a desired the treatment—or even seek it out themselves (e.g., buying an in-home display). For example, the Milton Hydro experiment sought to limit contact between participants in its different conditions (Schembri, 2008). Connecticut Light and Power's Plan-it Wise pilot randomly assigned participants without mentioning alternative conditions or allowing them to switch interventions if they did learn (Faruqui and Sergici, 2009a). See also the Ameren Illinois Power-Smart Pricing (PSP) pilot (Violette et al., 2010), BG&E Smart Energy Pricing Pilot (SEPP) (Faruqui and Sergici, 2009b), Pepco's PowerCents DC pilot (King, 2010), and Hydro Ottawa's Ontario Energy Board Smart Price Pilot (Strapp et al., 2007).

Attrition bias can be reduced with extrinsic incentives (e.g., completion bonuses) or intrinsic ones (e.g., stressing the value of complete data sets). Primary analyses should use an *intention-to-*

[3] http://www.consort-statement.org/home/

treat approach, where outcomes are estimated for all those enrolled in the trial according to the groups to which they were originally assigned, even if they do not complete the trial (Hollis and Campbell, 1999). At times, it may be possible to adjust for missing data using imputation methods (Ibrahim et al., 2005). See PG&E's Smart-Rate pilot (George et al., 2011) or Idaho Power's Energy Watch Pilot (Kline, 2007).

Acknowledgments

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 additional information about their studies. The views expressed are those of the authors.

Financial support was provided in part by the Carnegie Electricity Industry Center and 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 also 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.

Table A1
Studies included in meta-analysis.[a]

Study	Country	Interventions			Outcomes		*N*
		IHD	DP	Auto	Peak	Overall	
Hammerstrom et al. (2007)	US		✓	✓		✓	50
Sexton et al. (1987)	US	✓	✓		✓	✓	188
George et al. (2011)	US		✓	✓	✓	✓	1107
Frank (2008)	Can	✓				✓	424
Eiden (2009)	US	✓				✓	194
Seligman et al. (1978)	US	✓				✓	20
KEMA (2006)	US		✓	✓	✓	✓	100
McClelland and Cook (1979)	US	✓				✓	101
Dobson and Griffin (1992)	Can	✓				✓	99
Battalio et al. (1979)	US		✓			✓	93
Strapp et al. (2007)	Can		✓		✓		498
Violette et al. (2007)	US		✓	✓	✓	✓	448
Wolak (2007)	US		✓		✓	✓	123
Kline (2007)	US		✓		✓	✓	923
Puckett and Hennessy (2004)	US		✓	✓	✓	✓	195
Borenstein et al. (2002)	US		✓	✓	✓		242
King (2010)	US		✓	✓	✓		1160
Norton (2008)	US	✓				✓	2479
Faruqui and Sergici (2009b)	US	✓	✓	✓	✓		1375
Sulyma et al. (2008)	Can	✓	✓		✓	✓	837
Violette et al. (2010)	US		✓		✓	✓	5621
Faruqui and Sergici (2009a)	US	✓	✓	✓	✓	✓	1114
Sipe and Castor (2009)	US	✓				✓	586
Yun (2009)	US	✓				✓	8
Hutton et al. (1986)	US/Can	✓				✓	521
Parker et al. (2008)	US	✓				✓	18
Baladi et al. (1998)	US		✓		✓		411
Allen and Janda (2006)	US	✓				✓	60
RMI (2006)	US		✓	✓	✓		261
CRA (2005)	US		✓	✓	✓	✓	1815
Schembri (2008)	Can	✓	✓		✓	✓	108
SBC (2006, 2007)	US	✓	✓	✓		✓	7567
	27 (US)	17	21	12	18	26	

[a] IHD = in-home display, DP = dynamic pricing, Auto = automation, Peak = peak reduction, Overall = overall reduction, *N* = Total sample size.

Appendix A. Included studies

Studies included in meta-analysis are shown in Table A1.

Appendix B. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org.10.1016/j.enpol.2013.07.093.

References

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.

Allen, D., Janda, K., 2006. 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 American Council for an Energy-Efficient Economy Summer Study on Energy Efficiency in Buildings, pp. 7.1–7.12.

Altman, D., Bland, J., 1999a. Statistics notes: how to randomise. British Medical Journal 319 (7211), 703–704.

Altman, D., Bland, J., 1999b. Treatment allocation in controlled trials: why randomise?. British Medical Journal 318 (7192), 1209.

Baladi, M., Herriges, J., Sweeney, T., 1998. Residential response to voluntary time-of-use electricity rates. Resource and Energy Economics 20 (3), 225–244.

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.

Battalio, R., Kagel, J., Winkler, R., Winett, R., 1979. Residential electricity demand: an experimental study. The Review of Economics and Statistics 61 (2), 180–189.

Bland, J., Altman, D., 2000. Statistics notes: the odds ratio. British Medical Journal 320 (7247), 1468.

Borenstein, S., Jaske, M., Rosenfeld, A., 2002. Dynamic Pricing, Advanced Metering, and Demand Response in Electricity Markets. Technical Report, University of California Energy Institute UC Berkeley.

Brennan, R., Prediger, D., 1981. Coefficient kappa: some uses, misuses, and alternatives. Educational and Psychological Measurement 41 (3), 687–699.

Callahan, C., Hojat, M., Gonnella, J., 2007. Volunteer bias in medical education research: an empirical study of over three decades of longitudinal data. Medical Education 41 (8), 746–753.

Carleton, R., Sanders, C., Burack, W., 1960. Heparin administration after acute myocardial infarction. New England Journal of Medicine 263 (20), 1002–1005.

CRA, 2005. Impact evaluation of the California Statewide Pricing Pilot. Technical report, Charles River Associates (CRA).

Davis, A., Krishnamurti, T., 2013. The problems and solutions of predicting participation in energy efficiency programs. Applied Energy 111, 277–287.

Deeks, J., Dinnes, J., D'Amico, R., Sowden, A., Sakarovitch, C., Song, F., Petticrew, M., Altman, D., 2003. Evaluating non-randomised intervention studies. Health Technology Assessment 7 (27), 1–179.

Dobson, J., Griffin, J., 1992. Conservation effect of immediate electricity cost feedback on residential consumption behavior. In: Proceedings of the 7th American Council for an Energy-Efficient Economy Summer Study on Energy Efficiency in Buildings, pp. 10.33–10.35.

Eiden, J., 2009. Investigation into the Effects of Real-time, In-home Feedback to Conserve Energy in Residential Applications. Masters Thesis, University of Nebraska–Lincoln, unpublished.

Faruqui, A., Hledik, R., George, S., Bode, J., Mangasarian, P., Rohmund, I., Wikler, G., Ghosh, D., Yoshida, S., 2009. A National Assessment of Demand Response Potential. Technical Report, Government Printing Office, Washington, DC.

Faruqui, A., Sergici, 2009a. Connecticut Light and Power (CL&P) Impact Evaluation of CL&P's Plan-it-Wise energy program. Technical Report, Connecticut Light and Power and The Brattle Group.

Faruqui, A., Sergici, S., 2009b. BG&E Smart Energy Pricing Pilot (SEPP). Technical Report, Baltimore Gas and Electric Company and The Brattle Group.

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

Frank, S., 2008. Hydro One Networks Inc. Time-of-use Pricing Pilot Project Results. Technical Report, Hydro One Networks Inc.

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

George, S., Bode, J., Hartmann, B., 2011. PG&E Load Impact Evaluation of Pacific Gas and Electric Company's Time-based Pricing Tariffs. Technical Report, Freeman, Sullivan & Co. and Pacific Gas and Electric Company.

Grady, D., Herrington, D., Bittner, V., Blumenthal, R., Davidson, M., Hlatky, M., Hsia, J., Hulley, S., Herd, A., Khan, S., Newby, L., Waters, D., Vittinghoff, E., Wenger, N., 2002. Cardiovascular disease outcomes during 6.8 years of hormone therapy. Journal of the American Medical Association 288 (1), 49–57.

Hammerstrom, D., Ambrosio, R., Carlon, T., DeSteese, J., Horst, G., Kajfasz, R., Kiesling, L., Michie, P., Pratt, R., Yao, M., Brous, J., Chassin, D., Guttromson, R., Jarvegren, O., Katipamula, S., Le, N., Oliver, T., Thompson, S., 2007. Pacific Northwest Gridwise Testbed Demonstration Projects. Part I. Olympic Peninsula Project. Technical Report, Pacific Northwest National Laboratory.

Hartling, L., Bond, K., Vandermeer, B., Seida, J., Dryden, D., Rowe, B., 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.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–161.

Hedges, L., Vevea, J., 1998. Fixed- and random-effects models in meta-analysis. Psychological Methods 3 (4), 486–504.

Higgins, J., Altman, D., Sterne, J., 2011. Assessing risk of bias in included studies. In: Higgins, J., Green, S., (Eds.), Cochrane Handbook for Systematic Reviews of Interventions (5.1.0). Cochrane Reviews (Chapter 8). Available from 〈www.cochrane-handbook.org〉.

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

Hutton, R., Mauser, G., Filiatrault, P., Ahtola, O., 1986. Effects of cost-related feedback on consumer knowledge and consumption behavior: a field experimental approach. Journal of Consumer Research 13 (3), 327–336.

Ibrahim, J., Chen, M., Lipsitz, S., Herring, A., 2005. Missing-data methods for generalized linear models. Journal of the American Statistical Association 100 (469), 332–346.

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 (3), 777–779.

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

Jüni, P., Altman, D., Egger, M., 2001. Systematic reviews in health care: assessing the quality of controlled clinical trials. British Medical Journal 323 (7303), 42–46.

Karlowski, T., Chalmers, T., Frenkel, L., Kapikian, A., Lewis, T., Lynch, J., 1975. Ascorbic acid for the common cold. Journal of the American Medical Association 231 (10), 1038–1042.

KEMA 2006. SDG&E Smart Thermostat Program Impact Evaluation. Technical Report, San Diego Gas and Electric Company and KEMA.

King, C., 2010. PowerCents DC Program Final Report. Technical report, Potomac Electric Power Company and eMeter Strategic Consulting.

Kjaergard, L., Villumsen, J., Gluud, C., 2001. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Annals of Internal Medicine 135 (11), 982–989.

Kline, B., 2007. Idaho Power 2006 Analysis of the Residential Time-of-day and Energy Watch pilot Programs: Final Report. Technical Report, Idaho Power Company and RLW Analytics.

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

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

Moher, D., Hopewell, S., Schulz, K., Montori, V., Gøtzsche, P., Devereaux, P., Elbourne, D., Egger, M., Altman, D., 2010. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Journal of Clinical Epidemiology 63, e1–e37.

Moher, D., Pham, B., Jones, A., Cook, D., Jadad, A., Moher, M., Tugwell, P., Klassen, T., 1998. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? The Lancet 352 (9128), 609–613.

Neenan, B., Robinson, J., 2010. Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols. Technical Report, Electric Power Research Institute.

Norton, B., 2008. PowerCost Monitor Pilot Program Evaluation. Technical Report, National Grid, NSTAR Electric, Western Massachusetts Electric Company, and Opinion Dynamics Corporation.

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

Parker, D., Hoak, D., Cummings, J., 2008. Pilot Evaluation of Energy Savings from Residential Energy Demand Feedback Devices. Technical Report, Florida Solar Energy Center and the US Department of Energy.

Parsons, H., 1974. What happened at Hawthorne? Science 183 (4128), 922–932.

Petitti, D., 2004. Hormone replacement therapy and coronary heart disease: four lessons. International Journal of Epidemiology 33 (3), 461–463.

Puckett, C., Hennessy, T., 2004. AmerenUE Residential TOU Pilot Study Load Research Analysis: First Look Results. Technical Report, RLW Analytics and AmerenUE.

Rosenthal, R., Rosnow, R., 1975. The Volunteer Subject. John Wiley & Sons, Hoboken, NJ.

RMI, 2006. Automated Demand Response System Pilot, Final Report. Technical report, Pacific Gas & Electric, Southern California Edison, San Diego Gas & Electric, and the Rocky Mountain Institute (RMI).

SBC, 2006. Evaluation of the 2005 Energy-Smart Pricing Plan-Final Report. Technical Report, Community Energy Cooperative and Summit Blue Consulting (SBC).

SBC, 2007. Evaluation of the 2006 Energy-Smart Pricing Plan-Final Report. Technical Report, Community Energy Cooperative and Summit Blue Consulting (SBC).

Schembri, J., 2008. The Influence of Home Energy Management Systems on the Behaviours of Residential Electricity Consumers: An Ontario, Canada Case Study. Masters Thesis, University of Waterloo, unpublished.

Schulz, K., 1995a. Subverting randomization in controlled trials. Journal of the American Medical Association 274 (18), 1456–1458.

Schulz, K., 1995b. Unbiased research and the human spirit: the challenges of randomized controlled trials. Canadian Medical Association Journal 153 (6), 783–786.

Schulz, K., Chalmers, I., Hayes, R., Altman, D., 1995. Empirical evidence of bias. Journal of the American Medical Association 273 (5), 408–412.

Schulz, K., Grimes, D., 2002. Generation of allocation sequences in randomised trials: chance, not choice. The Lancet 359 (9305), 515–519.

Seligman, C., Darley, J., Becker, L., 1978. Behavioral approaches to residential energy conservation. Energy and Buildings 1 (3), 325–337.

Sexton, R., Johnson, N., Konakayama, A., 1987. Consumer response to continuous-display electricity-use monitors in a time-of-use pricing experiment. Journal of Consumer Research 14 (1), 55–62.

Shun-Shin, M., Thompson, M., Heneghan, C., Perera, R., Harnden, A., Mant, D., 2009. Neuraminidase inhibitors for treatment and prophylaxis of influenza in children: systematic review and meta-analysis of randomised controlled trials. British Medical Journal 339, b1372–b1381.

Sipe, B., Castor, S., 2009. The net impact of home energy feedback devices. In: 2009 Energy Program Evaluation Conference, Portland, pp. 341–351.

Spees, K., Lave, L., 2007. Impacts of responsive load in PJM load shifting and real time pricing. Carnegie Mellon Electricity Industry Center Working Paper CEIC-07-02, January 2007.

Spiegelhalter, D., Best, N., 2003. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. Statistics in Medicine 22 (23), 3687–3709.

Strapp, J., King, C., Talbott, S., 2007. Ontario Energy Board Smart Price Pilot final report. Technical Report, Ontario Energy Board and IBM Global Business Services and eMeter Strategic Consulting.

Stukel, T., Fisher, E., Wennberg, D., Alter, D., Gottlieb, D., Vermeulen, M., 2007. Analysis of observational studies in the presence of treatment selection bias. Journal of the American Medical Association 297 (3), 278–285.

Sulyma, I., Tiedemann, K., Pedersen, M., Rebman, M., Yu, M., Power Smart, BC Hydro. 2008. Experimental evidence: a residential time of use pilot. In: American Council for an Energy-Efficient Economy Summer Study on Energy Efficiency in Buildings, pp. 292–304.

Thompson, S., Higgins, J., 2002. How should meta-regression analyses be undertaken and interpreted? Statistics in Medicine 21 (11), 1559–1573.

Todd, A., Stuart, E., Schiller, S., Goldman, C., 2012. Evaluation, measurement, and verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations. Technical Report, Lawrence Berkeley National Laboratory.

Tune, G., 1964. Response preferences: a review of some relevant literature. Psychological Bulletin 61 (4), 286–302.

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

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

Violette, D., Erickson, J., Klos, M., 2007. Final Report for the MyPower Pricing Segments Evaluation. Technical Report, Public Service Electric and Gas Company and Summit Blue Consulting.

Violette, D., Provencher, B., Klos, M., Freeman, R., Steele-Mosey, P., Clark, D., Klos, D., 2010. Ameren Power Smart Pricing 2009 Annual Report. Technical Report, Ameren Illinois Utilities and Summit Blue Consulting.

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

Wolak, F., 2007. Residential Customer Response to Real-Time Pricing: The Anaheim Critical Peak Pricing Experiment. Stanford University, unpublished manuscript.

Wooldridge, J., 2002. Econometric Analysis of Cross Section and Panel Data. MIT press, Cambridge, MA.

Yun, T., 2009. 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, ACM, pp. 4417–4422.