WASHINGTON UTILITIES AND TRANSPORTATION COMMISSION STAFF
RESPONSE TO DATA REQUEST

DATE PREPARED:	December 21, 2011	WITNESS:	Deborah J. Reynolds
DOCKETS:	UE-111048/UG-111049	RESPONDER:	Deborah J. Reynolds
REQUESTER:	Puget Sound Energy	TELEPHONE:	(360) 664-1255

PSE Data Request No. 025 to WUTC Staff:
RE: Deborah J. Reynolds, Exhibit No. ___(DJR-1T), page 32, line 15

Please explain fully the criteria Commission Staff would use to evaluate whether a sufficient level of analysis was performed to meet a "known and measureable" standard. As part of this response, please fully explain what it means by a "statistically significant post-installation analysis."

RESPONSE:

Because PSE has provided no analysis of its estimates of energy savings in its direct case, it is difficult for Staff to respond to this data request. However, as a general rule, the "Analysis of PSE's Pilot Energy Conservation Project: 'Home Energy Reports'" prepared by Annika Todd, Steven Schiller and Charles Goldman in response to Staff's request for assistance from Lawrence Berkeley National Laboratory describes statistically significant post-implementation analysis, which in this case is an appropriate substitute for post-installation analysis. This document was circulated to PSE's Conservation Resources Advisory Group (CRAG) on October 17, 2011. See Attachment.

Analysis of PSE's Pilot Energy Conservation Project: "Home Energy Reports"

Lawrence Berkeley National Laboratory Technical Memo
Annika Todd, Steven Schiller, Charles Goldman
October 17, 2011

Executive Summary

Overall, with respect to evaluation of energy savings, the method of program implementation and analysis for Puget Sound Energy's Home Energy Reports (HER) program was excellent and the estimates of energy savings are valid (assuming that the data were valid and that the calculations were mechanically correct). However, LBNL is in agreement with KEMA's "20 Month Impact Evaluation[1]" that the results are only applicable to the study duration (20 months) and the study population (households in King County that use more than 80MBtus and are single family homes, among other restrictions). While the analysis methods used in this pilot are very robust, the savings estimates cannot be applied directly to a full-scale rollout of the program: for the currently defined study population, a control group that does not receive HERs should be maintained, and for a different population (such as low energy users) a new control group should be established in order to correctly estimate savings.

LBNL Review of Method, Analysis, and Results

- The evaluation study design for the HER pilot program utilized a randomized controlled experiment with an opt-out design, which is the *best feasible method* of inferring that a program caused energy savings. With this method, any difference in energy use between the control and treatment groups can be attributed to the HER program. With other methods that are commonly used, it is likely that savings estimates are biased.

[1] KEMA, 2010. *Puget Sound Energy's Home Energy Reports Program: 20 Month Impact Evaluation.* Madison, WI.

- KEMA's "20 Month Impact Evaluation" (denoted KEMA's Evaluation for the remainder of this memo) presented two methods for estimating energy savings for the HER program: the "Pooled" method and the "Difference-of-Differences" method. The KEMA Evaluation used the numbers from the "Pooled" method. LBNL believes that this method may have biased estimates and definitely has erroneous confidence intervals that are too small. However, the "Difference-of-Differences" method produces unbiased, statistically significant estimates of energy savings with correctly calculated confidence intervals. Therefore, the numbers from this estimation, presented in the last two columns of Table C-1 in KEMA's Evaluation, should be used instead of the numbers presented throughout KEMA's Evaluation from the "Pooled" method. The amount of total savings over 20 months from these two models is almost identical, although the first 12 months and last 12 months differ slightly.
- Specifically, LBNL believes that Table 2 below (which is excerpted from Table C-1 in KEMA's Evaluation and reflects the "Differences-of-Differences" method) provides the most robust estimate of energy savings. Note that the 95% confidence intervals do not include zero, indicating that these results are statistically significant. Thus, these results provide strong evidence that there are actual energy savings from the HER pilot program. These savings estimates are not adjusted for weather, as discussed further below.

Table 2: Annualized Estimated Savings per Treatment Household[2]

	First 12 months (11/08-10/09)		All 20 months (11/08-6/10, annualized)		Last 12 months (7/09-6/10)	
Electric Savings	**183.2 kWh**	**1.65%**	**204.5 kWh**	**1.84%**	**225.4 kWh**	**2.03%**
95% confidence interval	*±26.3 kWh*	*±0.24%*	*±28.3 kWh*	*0.26%*	*±33.6 kWh*	*0.30%*
Gas Savings	**10.7 Therms**	**1.11%**	**12.1 Therms**	**1.26%**	**13.4 Therms**	**1.40%**
95% confidence interval	*±1.8 Therms*	*0.19%*	*±1.9 Therms*	*0.20%*	*±2.3 Therms*	*0.24%*

- KEMA's March 7, 2011 memo entitled "Home Energy Report Evaluation – Analysis of PSE's EE Program Tracking Data" (denoted KEMA's Double Counting Memo for the remainder of this memo) provides a good analysis and estimates of the magnitude of the double counted savings for programs that were tracked. Specifically, see Table 3 and Table 4 below in section 6 for double counting numbers (excerpted from KEMA's Double Counting Memo Tables 2-5). Table 3 uses a "Time of Participation" method, while Table 4 uses a "Load Shape-Allocated" method. LBNL agrees with KEMA that both methods are sound and that PSE should use whichever method it believes is appropriate from an accounting perspective.

[2] Similarly, LBNL believes that the last two columns of Table C-2 and Table C-3 in KEMA's Evaluation titled "Differences-of-Differences" provides the most robust estimate of energy savings specific to monthly vs. quarterly reports.

- In meetings with PSE staff, they indicated that they were considering deducting the double counted savings from the HER program for ease of accounting purposes, but would not deduct these savings when considering the overall effectiveness of the HER program. LBNL agrees with PSE that for accounting purposes it may not matter which program receives the double counted savings. For considering the overall effectiveness of the HER program, LBNL recommends allocating the double counted savings such that the HER program receives between 50% and 100% and the other tracked program receives between 0% and 50% of the double counted savings, as discussed further below in section 6. Without any additional information, an intermediate case might be recommended, where the HER program receives 75% of the double counted savings and the other program receives 25%. Note that if the double counted savings are entirely given to the other program, this could create a perverse incentive for OPOWER to *not* direct customers to other programs.

LBNL Recommendations for Applicability of Results to Other Populations and Future Years

- LBNL agrees with KEMA's Evaluation that these estimates of energy savings are only valid for the study population, and should not be extrapolated outside of the study population to the greater PSE territory[3]. Specifically, because the population was restricted to King County and to households that use more than 80MBtus of energy (this energy restriction cut out approximately 12-15% of households after all other restrictions were applied), the savings estimates cannot be assumed to be the same for households outside of King County or for households that use less than 80MBtus of energy.
- LBNL agrees with KEMA's Evaluation that these estimates of energy savings are only valid for the study duration (20 months), and should not be extrapolated into future years; savings should be estimated each year using actual energy data for the past year from treatment and control groups. Specifically, these estimates are only applicable to the conditions that occurred during the study period, including weather, consumer energy costs, economic conditions, etc.

LBNL Recommendations Going Forward

- In the future, LBNL recommends that a randomly allocated control and treatment group should be maintained in order to allow unbiased estimates of energy savings each year. In practice, this means that HERs cannot ever be mailed to *every* household. However, it is

[3] The study population was restricted to single family, residential homes located in King County that use more than 80MBtu of energy per year, use both natural gas and electricity provided by PSE, do not use a solar PV system, have parcel data available from the county assessor, have a bill history that starts on or before Jan 1 2007, have 100 similarly sized homes within a two mile radius, and have automatic daily meter reads.

possible that the size of the control group could be reduced in future rollouts (50% of the study population is likely not needed). Analysis should be done to determine the smallest possible control group such that the estimates are likely to be statistically significant at 5%. If the control group size is reduced, then more people can be in the treatment group, and aggregate savings are likely to be higher[4].

- If the program is to be expanded to additional populations and additional counties, LBNL recommends that a new study population is defined and new control and treatment groups are randomly assigned from within the new study population. Again, an analysis should be done to determine the smallest possible control group so that as many households as possible can be placed in the treatment group.
- In future analysis, LBNL would recommend using either a difference-in-difference model defined in section 4.3.2 below as Model 1 (which is the same as the method used to produce the "Difference-of-Differences" results in the last two columns of Table C-1 in KEMA's Evaluation), or a fixed effects regression with standard errors clustered at the household level that is not normalized for a "typical" year's weather, defined in section 4.3.2 below as Model 2. Both models lead to unbiased estimates of energy savings with correctly calculated confidence intervals. Model 1 is a more simple analysis, while Model 2 may be slightly more precise in the sense that it may have slightly smaller confidence intervals.
- In the future LBNL recommends that KEMA or PSE continues to review program tracking databases to determine participation by customers in other PSE efficiency programs in order to calculate double counted savings for these tracked programs (using a method similar to that used in KEMA's Double Counting Memo).
- LBNL recommends that PSE consider conducting survey research of customers to assess the possible impact of programs that are not tracked, such as upstream programs that cannot be traced to a specific household. For example, the analysis described in Dougherty et al. 2011[5] used surveys of individual households in order to determine the types of measures (e.g., appliances, CFLs, and weatherization) that households in a Massachusetts HER program installed.

[4] Two good references for determining the optimal control and treatment sample sizes are (1) section four in: Duflo, E., R. Glennerster, and M. Kremer. 2007. "Using randomization in development economics research: A toolkit." *Handbook of development economics* 4: 3895-3962. http://economics.mit.edu/files/806; and (2) section 4 Protocol 5 in: "Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols." EPRI, Palo Alto, CA: 2010. 1020855.

[5] Doughterty, A., Dwelley, A., Henschel, R. and Hastings, R. *Moving Beyond Econometrics to Examine the Behavioral Changes behind Impacts.* IEPEC Conference Paper.

1 Introduction and Objective

LBNL was asked by the Washington Utilities and Transportation Commission (UTC) staff to provide an independent analytical review and critique of an emerging residential energy conservation behavior-change based program sponsored by Puget Sound Energy, called the Home Energy Report (HER) program. LBNL was also asked to provide a recommendation to accept, reject, or partially accept estimates of the energy savings attributable to the HER program presented in KEMA's "Puget Sound Energy's Home Energy Reports Program: 20 Month Impact Evaluation," as well as to make recommendations for the program and the analysis going forward.

This technical memo provides LBNL's review and assessment of KEMA's Evaluation of the HER program in response to the request by WA UTC staff. In this memo, LBNL reviews both the approach used by OPOWER and PSE to set up and implement the program as well as the analysis methodology used in KEMA's 20 Month Impact Evaluation. We also discuss implications and applicability of the results, and recommendations for continuing the HER program. This memo does not verify the validity of the data used or the calculations (i.e., we did not review the SAS code used in the analysis).

This memo addresses several specific issues that are important to the overall validity of the program's energy savings and is organized as follows. In section 2, we discuss issues related to the causal inference method (the method by which the savings can be causally attributed to the program; in this case, experimental design). In section 3, we discuss data related issues. In the following sections, we discuss analysis methodology issues (section 4), the external validity of the results (section 5), and double counting issues (section 6), and the final section provides recommendations for the future (section 7). Each section begins with a general discussion of a specific issue, including why the issue is important, the best practices for addressing the issue, and the implications of addressing the issue in different ways. Each section concludes with LBNL's assessment of the method by which the HER program setup and analysis addressed each issue and the implications of the method used. The main point of LBNL's assessment of the methods used is summarized for each issue with the label "LBNL Observation"; readers could quickly skim this memo reading only these observations, adding the surrounding discussion when necessary.

2 Causal Inference Method

General Discussion

The goal of the ongoing pilot project is to be able to infer whether or not Home Energy Reports (HERs) caused energy savings over a specific amount of time for a specific population. In order to determine whether or not energy savings were caused by the HERs, it is necessary to know the energy use of the specific population in the presence of HERs and the energy use without the HERs. Ideally, we

would be able to observe two parallel universes: one in which the customers received HERs, and one in which those exact same customers did not receive HERs, where the difference in energy use between the two is clearly the savings that was caused by the HERs. Because in reality a specific customer can either receive a HER or not, we can never observe the same customer in both situations for a specific time period.

Rather than observing the exact same customer in both situations, we can compare two groups of customers, one group who received HERs (the "treatment" group), and one group who did not (the "control" group). Then, any difference in energy use between the treatment and control groups comes from three sources: first, the treatment group received the HERs and the control group didn't; second, the people in the treatment group may be different than the people in the control group; and third, there is some inherent randomness. The key point is to try to minimize the differences between the people in the control and treatment groups, so that the difference in energy use can be attributed to the HER reports rather than differences between the people (statistics can then be used to determine whether the remaining differences are due to the HERs or to inherent randomness).

Randomized controlled trials (RCTs) are the best way to infer causality (if the HERs caused the observed changes in energy consumption). When customers in a defined population are randomly assigned to the treatment and control groups, the differences between the types of people in these two groups are minimized, and so any difference between the energy use of the treatment group and the control group can be causally attributed to the program. Sometimes "quasi-experimental" methods are used, in which customers in a program are compared to customers who are not in the program. The problem with this method is that these two groups may have different people in them (called "selection bias"). For example, customers who self-select into a program are obviously different types of people than those who don't, and programs are sometimes targeted to specific areas or specific demographic groups. In these examples, the difference in energy use between these two groups can be attributed both to the program and to pre-existing differences between people with these different characteristics. Even if all observable differences between the two groups are balanced or matched, there are always unobservable differences that are not matched (for example, the type of person who would sign up for a program).

Another method would be to compare the energy use of customers in a program to their own historical energy use; however, there are so many other factors that influence energy use (such as economic conditions, weather, political events, energy prices, other utility programs) that the difference between current energy use and past energy use that is attributable only to the program under consideration will always be very difficult to ascertain in a reliable and accurate manner.

Specific Implementation for the HER Program

LBNL Observation 1 *The causal inference method used in the HER Pilot Program was a randomized controlled experiment. This is the best method for causally attributing energy savings to HERs.*

The HER pilot was set up as a true randomized controlled experiment, where the treatment and control groups were randomly selected from the target population. This is the best method for inferring causality: any difference in energy use between the control group and treatment group can be attributed to the effect of the HERs (and to inherent randomness). Specifically, with randomization, the control and treatment groups should have equal proportions of both observable variables, such as income, energy use, and participation in other monitored utility programs (rebate programs), as well as unobservable variables, such as participation in non-monitored programs (CFL programs).

2.1 Randomization Design

General Discussion

There are several types of randomized controlled trial (RCT) designs, including mandatory RCT, RCT opt-out, RCT opt-in (which can be either recruit-and-delay or recruit-and-deny), crossover design, and factorial design. Mandatory RCT, in which people are randomly assigned to a control or treatment group with no option to opt-out of the treatment group, is good for determining the effect of a program but usually is not feasible. The next-best option is RCT opt-out, in which people are randomly assigned to a control or treatment group with an option to opt-out of the treatment group. In this case we can't tell how a program would affect those who opted out, but we probably don't want to force them to participate in any case. An opt-out design is better than an opt-in design because a much higher percentage of people tend to stay in an opt-out program than tend to opt-in to a voluntary program, and the types of people who tend to opt-in to a voluntary program are different than the types of people who don't opt-in to voluntary programs. RCT opt-in designs are only applicable to the type of customers who would opt-in to such a program; these customers may be a biased set of customers with different energy use characteristics than the general population.

Specific Implementation for the HER Program

LBNL Observation 2 *The experimental design used was a Randomized Controlled Trial with opt-out participants. LBNL believes that an opt-out design is the best feasible method for creating robust estimates of energy savings.*

The HER pilot project used an opt-out RCT design, where customers who were randomly allocated to the treatment group were mailed the HERs by default, and had to actively remove themselves from the

program if they no longer wished to receive the report[6]. This is the second best type of design, and we believe that in most cases it is the best design that is actually feasible.

2.2 Unit of Randomization

General Discussion

The treatment and control groups can be randomized over different units, where each unit is independently assigned to either group. The unit of randomization could be an individual, a household, a block, a town, etc. There are two issues to consider. First, for statistical significance, a large number of units are needed, and a smaller unit of randomization probably means more units (it would be hard to get 500 towns to participate in a randomized program). Second, the units should be large enough so that there are no spillover or externality effects between units. For example, while randomizing over individuals would result in more units than randomizing over households, individuals within a household can be expected to significantly influence each other's behavior, resulting in spillover outside of the unit of randomization. This can severely bias results, because then it becomes unclear who is "treated" and who is not, if household members are sharing information with each other.

Specific Implementation for the HER Program

LBNL Observation 3 *The unit of randomization was the household level. LBNL believes that this is the best unit of randomization for the HER program.*

LBNL Observation 4 *Spillover effects from one unit of randomization to another (from household to household) could result in biased estimates; however, LBNL agrees with OPOWER and KEMA that these spillover effects are not expected.*

The HER pilot was randomized at the household level. LBNL believes that spillover effects between households are not expected for the following reason. Because the HER letters are specific to a household, even if a household in the treatment group who received a HER shared information with a neighbor in the control group who did not receive a HER, the neighbor would not know their own standing relative to others. It is possible the act of discussing the letter with the neighbor might get the neighbor thinking more about his own energy use and how his energy use compares with others, and cause the neighbor to save energy. However, even if this is the case, this would cause the energy use of the control group to decrease, which would mean that the savings estimates are biased downwards (i.e., this would mean that the true energy savings are higher than the estimated savings).

[6] Around 1.6% of recipients opted-out of the program in the first two years. As discussed further below, these customers were not removed from the analysis.

2.3 Study Population

General Discussion

The study population is the group of people from which the control and treatment units are randomly assigned. The study population may be a specific, targeted subset of the entire population of customers, as long as both the control group and the treatment group are randomly assigned from the specific subset (that is, the control group cannot be taken from a different subset than the treatment group). It is important to clearly define the study population, because the estimated energy savings are only valid for the subset of customers in the study population without making strong assumptions about the program. For example, if the study population is the subset of customers that are high energy users, the energy savings results cannot be expected to be the same for low energy users[7].

There are two things that affect the definition of the study population: the type of experimental design, which implicitly restricts the study population; and the screening process, which explicitly restricts the study population. With opt-in designs, the study population is restricted to the type of people that would opt-in to the program. With randomized encouragement designs, the study population is restricted to a subset called "compliant" customers. With opt-out designs, there are two cases: if those who opt-out are *not* included in the analysis, the study population is restricted to the type of people who don't opt-out; if those who opt-out *are* included in the analysis, the study population is unrestricted by the experimental design (although it is still restricted by the screening process). Opt-out designs therefore are the most desirable because they do not restrict the study population for which the results of the program are valid.

The screening process also restricts the study population. Often the screening process restricts the study population to specific geographies (zip codes or service areas), specific demographics (low income, medical needs, elderly), specific customer characteristics (high energy users, dual fuel use, length of customer bill history), specific data requirements (census information is available, smart meter installed), and other restrictions. The choice of how to restrict and screen the study population is important. On one side, restricting the population means that the study's result can't be extrapolated outside that specific group. On the other side, it may be the case that the program works better for a certain subset of the population, and in this case it is more cost effective to limit the study population to this subset. Another reason to restrict the study population is for statistical precision; the more similar the households in the study group, the lower the variation in energy use, and the more precise the estimates become.

[7] In this case, one could make the assumption that low energy users would react to the HERs in the same way that high energy users would react, but this is a fairly strong assumption that is likely not true. On the other hand, if the study population were defined with a factor that is irrelevant to how people are likely to react to HERs, such as choosing the study population to be only those with blue eyes, then it is a weaker assumption to assume that people with brown eyes react in the same way as people with blue eyes (although it is still an assumption).

Specific Implementation for the HER Program

LBNL Observation 5 *The study population was restricted to customers that meet the following criteria: use dual fuel, are a single family residential home, are located in King County, use more than 80MBtu of energy per year, do not utilize a solar PV system, have an address that is available with parcel data from the county assessor, have a bill history that starts on or before Jan 1 2007, have 100 similar sized homes within a two mile radius, have automatic daily meter reads, and are not in the 98006 zip code. Because the experimental design is opt-out and the customers who opt-out are included in the analysis (as discussed further below), the study population is not further restricted by the type of experimental design.*

The HER pilot used an opt-out experimental design and included customers who opted-out in their analysis, which is the most conservative approach that places no implicit restrictions on the study population. The explicit screening process does restrict the study population to homes as described above. Some of the restrictions are for data purposes (bill history, parcel data), and some are to reduce variation in type of customer (excluding those who have a solar PV system). LBNL believes that all of these are valid restrictions, although as noted, the energy savings calculated for this pilot program should not be extrapolated outside of this defined, restricted study population.

2.4 Study Duration

General Discussion

The study duration is the length of time that the original, randomly allocated treatment group receives the program and the control group does not receive the program. It does not include any time during which baseline data is collected. The estimates of energy savings due to the program are only valid for the study duration, and cannot be extrapolated outside of the study duration to future years without making strong assumptions about the program.[8]

Specific Implementation for the HER Program

LBNL Observation 6 *The study duration evaluated in the KEMA Evaluation is 20 months.*

The HER pilot program is being evaluated over a study duration of 20 months, but the study will continue as originally designed with a treatment and control group into the future. A three year evaluation is planned.

[8] As discussed below, there are many reasons why we might expect the energy savings from the HER program to increase, decrease, or stay the same. So far, the KEMA report has shown an upswing in savings in the second year relative to the first year.

2.5 Stratification

General Discussion

Sometimes programs employ a stratified sampling method when restricting the study population or when randomizing units into treatment and control groups. This is done to make sure that a sub-population of interest is represented by enough units to be able to make statistical conclusions about the program effectiveness for that sub-population; however, it requires a specific type of analysis.

Specific Implementation for the HER Program

LBNL Observation 7 *Stratification was not performed in the selection of the study population or in the randomization of households into treatment and control groups.*

Because stratification was not used, it does not need to be corrected for in the analysis.

3 Data

This section describes data collection, data cleaning, and data sampling methods.

3.1 Data Collection

Data that are appropriate for the program and the type of analysis desired should be collected.

LBNL Observation 8 *LBNL believes that data appropriate for the type of analyses performed were collected.*

For the HER program, the collected data include: household energy usage data, frequency of report delivery, household square footage and other household characteristic data. Household usage data were collected by automated CellNet meters for each home included in the participant and control groups, and the data were gathered on daily intervals. County assessor data were used to identify home values, household square footage, and identify neighboring homes.

3.2 Data Cleaning

The way in which data are cleaned (removing outliers or missing observations) can have a relatively large impact on the estimates in the analysis. There should be a clear methodology for cleaning data that is based on knowledge of the industry or of the data collection process.

LBNL Observation 9 *There was a clear methodology for cleaning the data: electric reads greater than 300kWh per day and less than 2 kWh per day were excluded from the sample. Gas reads greater than 100Therms per day and less than 0Therms per day were*

also excluded from the sample. Data for households that did not have usable zip codes were excluded. LBNL agrees that this methodology for cleaning data is appropriate.

Data for households that closed accounts or opted-out of the program are discussed below.

3.3 Data Sampling

If the entire dataset is not used in the analysis for any reason, it is important to ensure that the sample of data used is a typical sample that is not biased in any way.

LBNL Observation 10 *All of the data were used; no data sampling took place. LBNL believes that this is the best way to create robust estimates of energy savings.*

4 Analysis Method

This section discusses the method used to analyze the dataset.

4.1 Balanced Randomization Check

General Discussion

If a population is large enough and is randomly assigned to a treatment and control group, then in theory the treatment group should have the same distribution of household characteristics as the control group. In practice, it is a good idea to check to make sure that this is true. Of course, only the observable characteristics of households can be tested.

Specific Implementation for the HER Program

LBNL Observation 11 *The treatment group was not found to be statistically significantly different than the control group when considering multiple household characteristics such as energy use, age of house, income, number of occupants, number of rooms, square feet, and whether the home is owned or rented, as shown below in Table 1. LBNL agrees with KEMA that this is sufficient evidence that the randomization was balanced and that therefore estimates of energy savings were not biased by differences between the two groups.*

KEMA tested each of the household characteristics listed below in Table 1 below to determine whether the mean of each characteristic was statistically different between the control and treatment group. For example, the mean electricity use in July 2007 was 853.3 for the treatment group, and was 854.8 for the control group. KEMA tested whether these two numbers were statistically significantly different than each other, and found that they were not, indicating that the control and treatment groups were not significantly different with respect to mean electricity use in July 2007. Specifically, if

the p-value, which is the number in the last column labeled "Pr > |t|" is less than 0.0012 (which is 0.05 level of significance divided by the number of tests, 41), then there should be some concern that the treatment and control groups are significantly different for that specific characteristic. For July 2007 mean energy use, the p-value is 0.6472, well above this cutoff. KEMA repeated this analysis for each of the characteristics listed, and found that every single one of the p-values is much higher than the cutoff of 0.0012; none of them is below 0.15, indicating that there is strong evidence that the HER program treatment and control groups are balanced along characteristics that are observable. There is always some risk that the unobservable characteristics are imbalanced and could cause bias in results, but we believe that this risk is very small because of the large-scale randomized controlled design of the study.

Table 1: Test of Balanced Sample (Reproduced from Table A-1 in KEMA's Evaluation)

Testing for a Balance Treatment/Control Sample, Individual Characteristic T-Tests

	Treatment			Control				
Characteristic	Count	Mean	SE	Count	Mean	SE	Difference	Pr> |t|
elecuse01JUL07	31,618	853.3	2.4657	40,006	854.8	2.2023	1.5136	0.6472
elecuse01AUG07	31,618	823.3	2.2955	40,006	823.6	2.0527	0.3203	0.9172
elecuse01SEP07	31,618	818.4	2.1534	40,006	820.1	1.927	1.7035	0.5558
elecuse01OCT07	31,618	920	2.3835	40,006	920.1	2.1152	0.1114	0.9721
elecuse01NOV07	31,618	998.1	2.6461	40,006	997.9	2.3092	-0.1528	0.9652
elecuse01DEC07	31,618	1217.8	3.3869	40,006	1218.1	2.9601	0.2409	0.9572
elecuse01JAN08	31,618	1105.8	3.0973	40,006	1103.8	2.6998	-2.0404	0.6187
elecuse01FEB08	31,618	947.2	2.6114	40,006	946.1	2.2945	-1.0043	0.7723
elecuse01MAR08	31,618	979.5	2.6819	40,006	980.5	2.3549	1.0055	0.7778
elecuse01APR08	31,618	877	2.3715	40,006	878.6	2.1034	1.5232	0.6308
elecuse01MAY08	31,618	838.1	2.2139	40,006	839.1	1.9748	1.0093	0.7338
elecuse01JUN08	31,618	810.7	2.169	40,006	812.5	1.9421	1.744	0.5495
gasuse01JUL07	31,619	18.931	0.093	40,007	18.9908	0.0848	0.0598	0.6358
gasuse01AUG07	31,619	20.0447	0.1074	40,007	20.0577	0.0965	0.0129	0.9287
gasuse01SEP07	31,619	32.4092	0.1128	40,007	32.4774	0.0954	0.0682	0.6426
gasuse01OCT07	31,619	76.1233	0.1676	40,007	76.1525	0.1481	0.0292	0.8959
gasuse01NOV07	31,619	110.7	0.2154	40,007	110.8	0.1898	0.0586	0.838
gasuse01DEC07	31,619	143.8	0.2686	40,007	143.9	0.2382	0.0627	0.8613
gasuse01JAN08	31,619	157.4	0.2879	40,007	157.4	0.2542	-0.0533	0.8895
gasuse01FEB08	31,619	114.7	0.2178	40,007	114.5	0.1915	-0.1657	0.5673
gasuse01MAR08	31,619	119.3	0.2304	40,007	119.4	0.2036	0.072	0.8146
gasuse01APR08	31,619	92.2053	0.189	40,007	92.2316	0.1674	0.0263	0.917
gasuse01MAY08	31,619	50.0173	0.1288	40,007	49.9791	0.112	-0.0383	0.822
gasuse01JUN08	31,619	41.1993	0.1248	40,007	41.1959	0.1091	-0.00343	0.9835
age	31,620	30.9307	0.0887	40,007	30.9408	0.0797	0.0101	0.9325
bedrooms	31,583	3.5499	0.00404	39,941	3.5449	0.0036	-0.00496	0.3595
bathrooms	31,620	2.2814	0.00329	40,007	2.2842	0.00293	0.00278	0.5281
fireplace	31,620	0.9569	0.00114	40,007	0.9549	0.00104	-0.00199	0.1975
house_value	31,614	347022	956.6	40,003	348235	869.5	1213.5	0.3491
income1	31,620	0.013	0.000636	40,007	0.012	0.000544	-0.00097	0.2452
income2	31,620	0.00794	0.000499	40,007	0.00787	0.000442	-0.00006	0.923
income3	31,620	0.0165	0.000716	40,007	0.0162	0.000631	-0.00028	0.7692
income4	31,620	0.0252	0.000881	40,007	0.0235	0.000758	-0.00163	0.1597
income5	31,620	0.0307	0.00097	40,007	0.0307	0.000862	-0.00001	0.9915
income6	31,620	0.1087	0.00175	40,007	0.1064	0.00154	-0.00228	0.3269
income7	31,620	0.1254	0.00186	40,007	0.1248	0.00165	-0.00062	0.8042
income8	31,620	0.1267	0.00187	40,007	0.1254	0.00166	-0.00131	0.5987
income9	31,620	0.4222	0.00278	40,007	0.4261	0.00247	0.0039	0.2944
num_occ	27,706	2.2168	0.00638	34,924	2.2287	0.00573	0.0118	0.1674
owned	27,706	0.9749	0.00094	34,924	0.9751	0.000834	0.000238	0.8495
sqft	31,620	2150.8	3.5589	40,007	2151.9	3.191	1.1429	0.8112

4.2 Attrition

General Discussion

Several types of attrition can happen throughout the duration of a program. People can opt-out of a program but still generate data after they've opted-out, as is the case with utility programs (unless the

design is mandatory, in which case they can't opt-out); people can disqualify during the program (by installing a solar PV, for example); and people can exit the program in such a way that their data is no longer available (for example, utility customers who move or close their accounts).

For those who opt-out but data are still available, including these people in the analysis is the most conservative method. Excluding them could lead to biased estimates, as it is likely that people who opt-out of a program are doing so because the program isn't working for them. If they are excluded, then the study population is restricted further to the type of people who don't opt-out of the program, and estimates of energy savings are only valid for this population.

For those who exit the program in such a way that data are no longer available, it is probably the case that these people exited for a reason other than the program, and so most likely people exited in the same rate from the treatment and control groups. An analysis can be done comparing characteristics of those who exited to make sure that the treatment and control groups are balanced. If the groups are balanced, the best way to deal with these people is to exclude them and all data derived from them entirely.

Specific Implementation for the HER Program

LBNL Observation 12 *The energy data for households that opted-out (around 1.6% of households) are included in the analysis (energy data for these customers is still available after they opt-out). LBNL agrees with KEMA that this is the best choice: it is a conservative way to estimate savings, and the estimates of energy savings are applicable to the entire study population rather than only to the types of customers that do not opt-out.*

LBNL Observation 13 *Data for households that exited the treatment or control group due to account closure or moving such that energy data was no longer available (roughly 10%) were excluded entirely. KEMA's analysis found that the distribution of these households in the control and treatment groups was "approximately balanced". LBNL therefore agrees with KEMA that excluding these data most likely had no effect on the analysis.*

KEMA's Evaluation includes data from customers who opted-out. This means that the estimate of energy savings can be interpreted as the savings due to placing a customer in the treatment group in the HER program (regardless of whether they later opt-out or not).

Another choice would be to exclude customers who opt-out, in which case the interpretation would be the savings due to receiving a HER. This would most likely result in a higher estimate of energy savings because it excludes the types of people who went out of their way to opt-out of the HER program. However, this measure is not as useful from a policy perspective because it only measures the effect of the HER program on a specific sub-population (those who are not the type of people who opt-

out), rather than the effect of the HER program on the population for which the program was originally intended. LBNL therefore agrees with KEMA's choice of including the opt-out customers in their analysis.

4.3 Model Selection for Estimating Energy Savings

General Discussion

There are several different analysis methods and models to choose from. The goal of any of these analyses is to create an estimate of energy savings due to the program that is: (1) unbiased, so that it does not under- or over-estimate the energy savings; (2) is internally valid, meaning that it is valid and unbiased for the given study population and given study duration; and (3) is as precise as possible, meaning that the 95% confidence intervals and the standard errors are correctly estimated and are reasonable (more on this below). The next section discusses precision, and the following section presents two models, both of which lead to unbiased estimates of energy savings attributable to a program.

4.3.1 Precision, Confidence Intervals, and Standard Errors

Recall from the beginning of this memo that with a randomized controlled trial, any energy savings for the treatment group relative to the control group can be attributed to three sources: first, the treatment group received the program while the control group did not; second, people in the control group may be different than the people in the treatment group; and third, there is some inherent randomness. We are interested in the first of these sources, which is measuring the energy savings attributable to the program. The energy savings attributable to the second source is minimized by using a randomized controlled trial design (and verifying that the groups are balanced), and so we are left with trying to ascertain how much of the savings is attributable to the first source (i.e., attributable to the program), and how much is attributable to the third source (i.e., attributable to inherent randomness). This section discusses the third source.

Consider the following example. Suppose there are only two households. The energy use of one household is likely to be quite different from the energy use of the other household, because each household has some inherent randomness in the way that they consume energy: the houses may be different sizes, have different appliances, or have different attitudes towards energy; one house may have been on vacation or may have hosted a large event; and there may be many other random differences. Suppose further that one of these two households is labeled a treatment household and receives the HER program, and that the other household is labeled a control household, and does not receive the HER program. If the treatment household is found to have used 2% less energy compared to the control household, some of the difference in energy use may be due to the program, but it is likely that most of it is due to this inherent randomness in energy use. In this case, the estimate of 2% energy

savings due to the program is not very *precise* because the true energy savings could be much higher or much lower than 2%.

So what we are really interested in is a *point estimate* of the energy savings (2% in this example) together with a *95% confidence interval*, which is the interval within which we are 95% certain that the true energy savings lies. For this example, it may be that the 95% confidence interval is (-8%,12%), meaning that with 95% probability, the true energy savings due to the program is somewhere between negative 8% and 10%. The 95% confidence interval is based on the *standard error* of the point estimate, where a confidence interval is roughly the point estimate plus or minus two standard errors. Standard errors are a measure of the inherent randomness of the data being measured, and take into account both the randomness of each unit and the randomness of the total number of units being measured. If there are more units, the standard errors decrease, and if there is less individual randomness in each unit, the standard errors decrease. Here, we will say that an estimate is more *precise* if it has smaller standard errors and therefore a smaller confidence interval.

Now suppose that we have 100,000 households in each group. With this many people, the inherent randomness in each household's energy use tend to balance each other out, and so the inherent randomness of the 100,000 households as a whole is smaller. This means that the standard errors and the confidence interval are smaller than in the case with only two households, and the estimate of energy savings is more precise[9].

Statistical Significance

While it is informative to know the confidence interval around a point estimate, often a binary decision has to be made: to either accept the estimate of energy savings (and therefore attribute that savings to the program), or don't accept the estimate. It is therefore useful to have a rule to use. Convention among scientists is to say that if the 95% confidence interval does not include zero, the estimate is statistically significant at 5% and the estimate should therefore be accepted.

While increasing the requirement to being statistically significant at 1% (or equivalently, that the 99% confidence interval doesn't include zero) would lead to more certainty about the estimate, it also increases the risk that an estimate of energy savings is rejected when in fact there are true energy savings. In practice, a requirement of 1% statistical significance would mean that programs would have to increase the number of people in a control group in order to sufficiently reduce standard errors, leading to fewer people in the treatment group and therefore lower total energy savings: in effect, it would increase the cost of the program.

Clustered Standard Errors

Returning to the example in which there are only two households, now imagine that 100 months of energy data were collected for each household. One way of analyzing this data is to assume that for each household, each month's energy use is independent of any other month's energy use, so that

[9] This is known in probability theory as the law of large numbers.

recording 100 months of data for one household is the same as recording data for 100 households for one month each. This would mean that the standard errors would decrease, because there are now effectively 200 total households.

However, energy use for one household is clearly not independent across the months: if the household uses a small amount of energy in one month relative to others, perhaps because it is a small apartment, then they are likely to use a relatively small amount of energy in the following months (this is called *serial correlation*). Therefore, analyzing the data as if it is independent in each month (acting as though there are 200 total households when in fact there are only 2) leads to erroneous, misleadingly small standard errors and confidence intervals[10]. On the other hand, it must be true that 100 months of data for two households contains more information than one month of data for two households.

There are two easy solutions to this serial correlation problem. The first solution, which works well when there are more than around 50 units (households in this case), is to use *standard errors that are clustered at the unit of randomization* in an analysis that uses data for each unit over time (such as a fixed effects method, discussed below). This method is easily implemented in most statistical packages.[11] Clustering standard errors essentially estimates the degree of independence in the data for each household over time and incorporates that into the standard errors: it uses all of the extra information available from having the multiple months of data, but also doesn't assume that each month is a completely separate household.

The second solution is to use a difference-in-difference model where the data are aggregated (described in more detail below). Basically, if there are 100 months of data for each household, this method averages over those 100 months so that there is one number for that household (its average energy use over 100 months). In this case, since we are collapsing 100 months into one average month, we don't have to worry about the data points being serially correlated over time. On the other hand, this method doesn't take into account the extra information that may be available in having 100 months rather than one average month. Therefore the standard errors (and the confidence interval) from a difference-in-difference model may be slightly larger than those from a model that uses the information from all of the months, such as the fixed effects model with clustered standard errors (described below). Even if they are slightly larger, the standard errors from a difference-in-difference model are much better than standard errors from a model in which the standard errors are *not* clustered (which can be three or more times smaller than the true, clustered standard errors).

[10] For a description of this effect with an example in which standard errors more than double, see: M. Bertrand, E. Duflo, and S. Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?," *Quarterly Journal of Economics* 119, no. 1 (2004): 249-275. http://econ-www.mit.edu/files/750

[11] See, for example, footnote #24 on page 271 of the above, http://econ-www.mit.edu/files/750

4.3.2 Two Unbiased Models

Next we turn to two specific models. The first type of model, which is the easiest computationally (that is, it takes the least time for a statistical package such as Stata or SAS to run), is a difference-in-differences model where the data are aggregated in such a way that there are only four numbers: (1) energy use, averaged over all people in the control group over 12 months before the program started, denoted by E(control, 12 months before); (2) energy use, averaged over all people in the treatment group over 12 months before the program started, denoted by E(treatment, 12 months before); (3) energy use, averaged over all people in the control group over 12 months after the program started, denoted by E(control, 12 months after); and (4) energy use, averaged over all people in the treatment group over 12 months after the program started, denoted by E(treatment, 12 months after). The effect of the program over the 12 months since the program started is estimated by calculating how much the treatment group changed their energy use relative to how much the control group changed their energy use:

Model 1:

Savings= [E(treatment, 12 months before)-E(treatment, 12 months after)] - [E(control, 12 months before)-E(control, 12 months after)]

Standard errors are calculated using a t-test

Note that this method relies on the assumption that the program begins for every control and treatment household that is being analyzed at the same time. For example, if the treatment coincides with the billing cycle and billing cycles are different for different households, then Model 1 is not appropriate and should not be used; instead, Model 2 below should be used.

The time period can be changed to 15 months or 3 months or any amount of time, as long as it's the same for all four numbers. Standard errors are calculated using a t-test (although a regression method can also be used). This model is relatively intuitive, gives an unbiased estimate of energy savings, and is internally valid, but it may have lower precision than other models for two reasons: first, it doesn't control for the inherent variability in energy use in different times of the year; and second, because the energy use is averaged over the 12 months, it doesn't use all of the information provided by the energy use in each month.

The second type of model, a fixed effects model, is more precise because it controls for the inherent variation of energy use in different times of the year and uses the energy information in every month, but it is computationally more difficult. It includes what are called unit-specific fixed effects for each unit *i* (if the unit of randomization is a household, these are household-specific fixed effects for each household *i*), and time fixed effects, which could be daily fixed effects, month-of-year fixed effects (where there are twelve fixed effects, and the effect for January is measured for every January that occurs in the sample), month-of-sample fixed effects (where there are as many fixed effects as there are months in the sample, and the effect for January of one year is estimated separately from the effect for January of another year), or another type of time effect. Consider Model 2 below which has three

slightly different variants. In this model a unit is assumed to be a household, and energy use is assumed to be collected on a monthly basis:

Model 2:

*(2a) EnergyUse(i,t) = a(i)+G*Post(t) +B*Treatment(i,t)+error(i,t)*

*(2b) EnergyUse(i,t) = a(i)+g(month-of-year)+G*Post(t)+B*Treatment(i,t)+error(i,t)*

(2c)[12] *EnergyUse(i,t) = a(i)+g(month-of-sample)+B*Treatment(i,t)+error(i,t)*

Where EnergyUse(i,t) is household i's energy use during month t, a(i) is a household-specific fixed effect, g(month-of-sample) and g(month-of-year) are time fixed effects, Post(t) takes the value 1 in months after the treatment begins for all households and 0 otherwise, Treatment(i,t) takes the value 1 if customer i is being treated during time period t and takes the value 0 otherwise, error(i,t) is an error term, B is the coefficient of interest, and standard errors are clustered at the household level.

With any of the specifications in Model 2, a regression will give an estimate of B which can be interpreted as follows: a household that is in the treatment group saves B units of energy per month on average relative to a household in the control group (assuming B is negative). Because the design was a randomized controlled experimental design, we can *causally* assign this savings to the program, and so the interpretation becomes: the program *causes* B units of energy to be saved per month per household that was in the program on average (assuming B is negative). This estimate is an unbiased estimate of energy savings over the period that is being analyzed for the given study population.

Because both Model 2 and Model 1 are unbiased, they should both result in the same estimate of energy savings, but Model 2 may be slightly more precise than the Model 1; Model 2 will have a slightly smaller confidence interval and slightly smaller standard errors (assuming that the standard errors are clustered at the unit of randomization). If the standard errors are not clustered at the unit of randomization, Model 2 gives an unbiased estimate of energy savings, but it will report incorrect standard errors and confidence intervals that appear to be much smaller than they actually are.

While all of the variants in Model 2 (2a, 2b, and 2c) will give unbiased estimates of B, (2b) may be slightly more precise than (2a) but will be slightly more computationally arduous because it includes twelve extra dummy variables, and (2c) may be slightly more precise than (2b) but will include even more dummy variables.

Adding Extra Variables

There are two cases of models that add additional variables to Model 2. In the first case, only *control variables* are added in an attempt to increase the precision even more (by reducing the standard errors and the confidence intervals). These control variables could include weather variables, such as heating degree days and cooling degree days or average temperature, or any other variable that changes over time. Control variables enter into the equation in Model 2 as a coefficient times the variable, so for example, *EnergyUse(i,t) = a(i)+g(f(t))+B*Treatment(i,t)+C*HDD(i,t)+error(i,t)* where C is the coefficient

[12] If there are 12 or fewer months, then (2c) should be used rather than (2b), because in that case Post(t) is not identified.

and HDD(i,t) is the heating degree days for customer i in time period t. Adding extra control variables probably won't cause bias in the estimated savings as long as too many aren't added.

The second case is adding *interaction variables* that enter into the equation in Model 2 as a coefficient times some variable *times the treatment variable*, so: $EnergyUse(i,t) = a(i)+g(f(t))+B*Treatment(i,t)+D*HDD*Treatment(i,t)+error(i,t)$, where HDD*Treatment(i,t) is an interaction variable because it describes the interaction between the two multiplied variables. While this type of model can be used to answer interesting questions about the program (in this example, estimating the coefficient D might tell us that the program works especially well on hot days), if the assumptions made in the model are not correct (in this case, that each additional heating degree day increases the effect of the program in a linear way), it could bias the estimate that we are actually interested in, which is the basic estimate of energy savings. On the other hand, if the assumptions that the model makes by including those variables is absolutely correct, then it would give an estimate of the energy savings that is exactly the same as a model that doesn't include the interactions.

Therefore a model with interaction terms should only be estimated as an *additional* analysis in order to gain deeper understandings about the program, but should not be used to estimate the basic energy savings.

Weather Normalization

We will now discuss the addition of specific interaction variables: those that are intended to normalize energy savings by weather. If the purpose of the analysis is to create a predictive model in which the program's impact in future years can be calculated simply by plugging in the future years' conditions (typically, HDD and CDD), then it might be worth including various interaction variables and testing their functional form.

However, creating a predictive model is not the primary objective. We have energy data from the control and treatment groups and so we can estimate the *actual* savings that occurred in the past year or past two years. Estimating the *actual* savings is much more precise than plugging weather variables into a model that predicts savings. Perhaps in the future, when there are 10 years of data for multiple behavioral programs in multiple areas, a predictive model like this could be of use (although such a predictive model should also include other factors that impact energy such as economic conditions).

Specific Implementation for the HER Program

LBNL Observation 14 *The KEMA Evaluation presents the "Pooled Specification Model" (described in their report on pages B-3 through B-7) as their preferred method for calculating energy savings, and results from this model are used throughout their evaluation. However, LBNL recommends that this model is not used. It is a fixed effects model in the form of Model 2 above, but it (a) includes multiple interaction variables, potentially leading to biased estimates, and (b) does not cluster the standard errors at*

the unit of randomization (the household level), resulting in incorrect, misleadingly small confidence intervals.

LBNL Observation 15 *The KEMA Evaluation also presents the "Difference-of-Differences" model (described in their report on pages B-1 and B-2). LBNL agrees with KEMA that this model results in unbiased estimates of energy savings, with correctly calculated standard errors and confidence intervals. The results from this model provide strong evidence that the HER program resulted in actual savings. LBNL therefore recommends that the energy savings estimates from this model should be used.*

LBNL Observation 16 *The HER reports were mailed at the same time to every customer in the study population. LBNL therefore agrees with KEMA that the "Difference-of-Differences" model is well defined.*

LBNL Observation 17 *Specifically, LBNL believes that Table 2 below (which is excerpted from Table C-1 in KEMA's Evaluation and reflects the "Differences-of-Differences" method) provides the most robust estimate of energy savings. Note that the 95% confidence intervals do not include zero, indicating that these results are statistically significant. Thus, these results provide strong evidence that there are actual energy savings from the HER pilot program. These savings estimates are not adjusted for weather.*

Table 2: Annualized Estimated Savings per Treatment Household[13]

	First 12 months (11/08-10/09)		All 20 months (11/08-6/10, annualized)		Last 12 months (7/09-6/10)	
Electric Savings	**183.2 kWh**	**1.65%**	**204.5 kWh**	**1.84%**	**225.4 kWh**	**2.03%**
95% confidence interval	*±26.3 kWh*	*±0.24%*	*±28.3 kWh*	*0.26%*	*±33.6 kWh*	*0.30%*
Gas Savings	**10.7 Therms**	**1.11%**	**12.1 Therms**	**1.26%**	**13.4 Therms**	**1.40%**
95% confidence interval	*±1.8 Therms*	*0.19%*	*±1.9 Therms*	*0.20%*	*±2.3 Therms*	*0.24%*

Most of the estimates of energy savings cited in KEMA's report come from the "Pooled Specification Model", given by:

"Pooled Specification Model" from pages B-3 through B-7:

$$EnergyUse(i,t) = a(i)+g(month\text{-}of\text{-}sample)+B*Treatment(i,t)$$
$$+C1*HDD(i,t)+C2*CDD(i,t)+D1*HDD(i,t)*Treatment(i,t)+CDD(i,t)*Treatment(i,t) + error(i,t)$$

[13] Similarly, LBNL believes that the last two columns of Table C-2 and Table C-3 in KEMA's Evaluation titled "Differences-of-Differences" provides the most robust estimate of energy savings specific to monthly vs. quarterly reports.

Where EnergyUse(i,t) is household i's energy use during month t; a(i) is a household specific fixed effect; g(month-of-sample) is a time fixed effect; HDD(i,t) and CDD(i,t) are heating and cooling degree days, respectively; Treatment(i,t) takes the value 1 if customer i is being treated during time period t and takes the value 0 otherwise; and error(i,t) is an error term.

Standard errors are NOT clustered at the unit level.

Note that although the specification in the report includes additional variables, these variables are not identified and were actually excluded in KEMA's analysis, and so this represents the model that was actually estimated.

Notice two features of this "Pooled Specification Model": first, it includes both extra control variables, which are labeled above with coefficients C1 and C2, as well as extra interaction variables, which are labeled above with coefficients D1 and D2; second, the standard errors are not clustered. The extra interaction variables can lead to a biased estimate of energy savings, as discussed above, and standard errors that are not clustered can lead to erroneous, misleadingly small standard errors and confidence intervals. Therefore, while this model can provide some interesting insights into the HER program, such as whether HERs result in higher savings on hotter days, we believe that the basic estimation of total energy saved due to HERs should not be based on this model.

Instead, LBNL believes that the basic energy savings estimates due to the HER program should be based on the results from the "Difference-of-Differences" model described on page B-1 and B-2, which is the same as Model 1 above. The "Difference-of-Differences" approach results in unbiased estimates of energy savings, with correctly calculated standard errors and confidence intervals (these results are given in the last two columns of Table C-1, titled "Difference-of-Difference"). It is possible that the precision of the estimates from the "Difference-of-Difference" model may be slightly improved (that is, the standard errors and confidence intervals may be slightly reduced) by using Model 2 above to estimate energy savings (where *all* of the features of Model 2 are adhered to, including clustered standard errors and no additional interaction variables); however, this was not done in the current analysis.

The reason given in KEMA's Evaluation for including the extra interaction variables in the "Pooled Specification Model" was described to be (from page B-4 in their report): "the savings should be put on a typical year basis, so that savings do not reflect consumptions pattern from an evaluation timeframe defined by atypical weather." As discussed above; while this approach might be useful when trying to predict *future* energy savings, the quantity that we are interested in is the best estimate of *past* energy savings: the *actual* savings that occurred in the previous year (or previous 20 months). In other words, we are not interested in coming up with estimates of energy savings in a typical year, because likely there are no "typical" years with a program such as HER. Energy savings due to HERs could increase or decrease over time and so trying to predict energy savings for *typical* years in the *future* is probably unreliable and unrealistic. Instead, we are interested in coming up with estimates of energy savings that actually occurred in the previous year due to HERs, given the weather (and the economic climate and the current events, etc) that actually occurred: Model 1 or Model 2 above is the best way to do this.

LBNL Observation 18 *LBNL recommends that models that estimate energy savings for a "typical" year are not used; instead, estimates of actual energy savings based on data from previous years should be used. We do not believe that there is enough evidence to suggest that a HER program has a "typical" year of energy savings*[14].

Note also that the presence of a control group completely controls for all possible weather effects, including HDD and CDD and any other weather event (snowstorms, humidity, etc), as well as any other non-weather events that happen (the super bowl, a stock market crash, etc).

4.4 Robustness Checks

It is usually a good idea to check the robustness of a model by changing some of the assumptions, re-estimating the effect of interest, and then thinking about why the results might be different.

LBNL Observation 19 *The KEMA Evaluation presented results from two different models. Despite the issues discussed above, the estimates for energy savings with the two models were relatively close to each other. LBNL believes that this indicates that the energy savings estimates are robust. This provides further evidence that the HER program results in actual energy savings.*

5 External Validity: Applicability of Results to Other Populations and Future Years

This section discusses external validity, or the extrapolation of savings estimates outside the study duration to future years, and outside the study population to other populations.

General Discussion

In general, results cannot be extrapolated beyond the study duration or outside of the study population. That is, even if energy savings have been estimated in an unbiased way for one year for a subset of people, it does not mean that those same energy savings will appear in a second year or for a different subset of people. Many other changes occur over time that can influence energy use, and so assuming that a program works the same way in future years is a very strong assumption. Likewise, different people are likely to react to programs in different ways, and assuming that all people will react to the program in the same way is a very strong assumption.

Specific Implementation for the HER Program

[14] In fact, the estimates of energy savings appear to have *increased* from year one to year two; if this trend continues, creating a "typical" year's savings would severely underestimate the actual energy savings.

LBNL Observation 20 *LBNL agrees with KEMA's Evaluation that these estimates of energy savings are only valid for the study population, and should not be extrapolated outside of the study population to the greater PSE territory*[15]*. Specifically, because the population was restricted to King County and to households that use more than 80MBtus of energy (this energy restriction cut out approximately 12-15% of households after all other restrictions were applied), the savings estimates cannot be assumed to be the same for households outside of King County or for households that use less than 80MBtus of energy.*

LBNL Observation 21 *LBNL agrees with KEMA's Evaluation that these estimates of energy savings are only valid for the study duration (20 months), and should not be extrapolated into future years; savings should be estimated each year using actual energy data for the past year from treatment and control groups. Specifically, these estimates are only applicable to the conditions that occurred during the study period, including weather, consumer energy costs, economic conditions, etc.*

It may be the case that the effect of HERs increases over time, as customers become more conscious of their energy use and form energy conserving habits, or it may be the case that the effects decrease over time, as people become inured to receiving the letters. In either case, assuming that future savings are the same as past savings is risky and probably not true.

Similarly, results for customers that are not in the current study population (low energy users, multi-family homes, etc.) should not be expected to be the same as the results for the current study population.

6 Double Counting

6.1 Other Programs That Are Tracked

General Discussion

Consider the following example, while assuming that participation in other programs can be tracked for each household. In addition to the HER program, there is a CFL rebate program. People must enter their address to receive a CFL rebate, so it is known with certainty which households used the rebates, and specifically, whether each household that used a rebate was part of the HER treatment group or the HER control group. Suppose that in the HER control group, 50 households used a CFL rebate, and in the

[15] The study population was restricted to single family, residential homes located in King County that use more than 80MBtu of energy per year, use both natural gas and electricity provided by PSE, do not use a solar PV system, have parcel data available from the county assessor, have a bill history that starts on or before Jan 1 2007, have 100 similarly sized homes within a two mile radius, and have automatic daily meter reads.

HER treatment group, 75 households used a CFL rebate. While the HER program is experimentally designed, so that it has both a treatment group that is exposed to the HER program and a control group that is not exposed to the HER program, the CFL rebate program in effect only has a treatment group: all households are exposed to the CFL program because anyone can receive the CFL rebates. So we can never observe the number of households that would have bought CFLs in the absence of the CFL program:

		HER Program	
		Control (not exposed to HER program)	Treatment (exposed to HER program)
CFL Program	Treatment (exposed to CFL rebate program)	50	75
	Control (not exposed to CFL rebate program)	?	?

As discussed in KEMA's Double Counting Memo, savings may be double counted by both the HER program and other programs only if the savings from measure installations are higher among households in the treatment group than those in the control group. In this example, 25 CFLs are double counted by both the HER program and the CFL program; the 50 CFL rebates that are used in both the control and treatment groups are only counted by the CFL program.

Because households were randomly assigned to the treatment and control groups for the HER program, as discussed above in section 2, any difference between the two groups can be attributed to the HER program (or to random noise, which can be addressed through statistical tests). Therefore, in this example, the HER program *caused* 25 extra people to participate in the CFL program by using a rebate: the HER program is a *necessary condition* for those 25 rebates.

The question then becomes: was the CFL program also a necessary condition for those 25 rebates?. To answer this question, consider two extreme cases. In Case 1, The CFL program was not a necessary condition: the 25 extra households in the treatment group that used CFL rebates were motivated by the HERs to purchase a CFL, and would have bought them regardless of if there were a rebate or not (but since it was available, they used the rebate). In this case, if there were a control group that wasn't exposed to the CFL rebate program, we would see 25 more CFL rebates in the HER treatment group as compared to the HER control group:

Case 1: *HER Necessary, CFL Not Necessary: 100% of double counted savings to HER.*		HER Program	
		Control (not exposed to HER program)	Treatment (exposed to HER program)
CFL Program	Treatment (exposed to CFL rebate program)	50	75 (25 more)
	Control (not exposed to CFL rebate program)	20	45 (25 more)

In this case, clearly 100% of the double counted savings should go to the HER program. At the other extreme, in Case 2, the CFL program was also a necessary condition: the 25 extra households that used CFLs wouldn't have bought any CFLs without the CFL rebate. In this case, if there were a control group that wasn't exposed to the CFL rebate program, we would see 0 more CFL rebates in the HER treatment group as compared to the HER control group. In this case, because both programs were necessary conditions to get the extra 25 CFL rebates, we might want to split the double counted savings with 50% for each program:

Case 2: *HER Necessary, CFL Necessary: 50% of double counted savings to each.*		HER Program	
		Control (not exposed to HER program)	Treatment (exposed to HER program)
CFL Program	Treatment (exposed to CFL rebate program)	50	75 (25 more)
	Control (not exposed to CFL rebate program)	20	20 (0 more)

Because the CFL program doesn't have a control group, we can't tell which of these cases is correct. Without any additional information, we might choose an intermediate case, where the HER program receives 75% of the double counted savings and the tracked (CFL) program receives 25%.

Note that if the double counted savings are entirely given to the other, tracked program, this could create a perverse incentive for OPOWER to *not* direct customers to other programs.

Specific Implementation for the HER Program

LBNL Observation 22 *KEMA's Double Counting Memo provides a good analysis and estimates of the magnitude of the double counted savings for programs that were tracked. Specifically, see Table 3 and Table 4 below for double counting numbers (excerpted from KEMA's Double Counting Memo Tables 2-5). Table 3 uses a "Time of Participation" method, while Table 4 uses a "Load Shape-Allocated" method. LBNL*

agrees with KEMA that both methods are sound and that PSE should use whichever method it believes is appropriate from an accounting perspective.[16]

LBNL Observation 23 *In meetings with PSE staff, they indicated that they were considering deducting the double counted savings from the HER program for ease of accounting purposes, but would not deduct these savings when considering the overall effectiveness of the HER program. LBNL agrees with PSE that for accounting purposes it may not matter which program receives the double counted savings. For considering the overall effectiveness of the HER program, LBNL recommends allocating the double counted savings such that the HER program receives between 50% and 100% and the other tracked program receives between 0% and 50% of the double counted savings, as discussed further below in section 6. Without any additional information, an intermediate case might be recommended, where the HER program receives 75% of the double counted savings and the tracked (CFL) program receives 25%.*

Table 3: Double Counted Savings for Tracked Programs, Time of Participation Method

		Year 1 (11/08-10/09)	Year 2 (11/09-10/10)	Both Years (11/08-10/10)
Electric	Test-Control (total double counted kWh)	93,711	5,736	99,447
	Double counted kWh per person in treatment (Divided by 27,094)	3.46	0.21	3.67
Gas	Test-Control (total double counted kWh)	34,703	45,810	80,512
	Double counted kWh per person in treatment (Divided by 27,094)	1.28	1.69	2.97

[16] The method used by KEMA can be done in either an aggregated form, as was presented in their Double Counting Memo, or can be done for each program separately; either method will result in the same estimate of double counted savings.

Table 4: Double Counted Savings for Tracked Programs, Load Shape-Allocated Method

		Year 1 (11/08-10/09)	Year 2 (11/09-10/10)	Both Years (11/08-10/10)
Electric	Test-Control (total double counted kWh)	25,580	76,605	102,185
	Double counted kWh per person in treatment (Divided by 27,094)	0.94	2.83	3.77
Gas	Test-Control (total double counted kWh)	8,424	45,345	53,768
	Double counted kWh per person in treatment (Divided by 27,094)	0.31	1.67	1.98

6.2 Other Programs That Are Not Tracked

General Discussion

While the example above used a CFL program as an example of a tracked program, in reality, CFL programs are usually targeted upstream and can't be tracked to a specific household. For programs that can't be tracked, in addition to the uncertainty about whether 50% or 100% of the double counted savings should go to the HER program, there is also uncertainty as to the actual magnitude of the double counted savings.

One method of estimating the magnitude of double counted savings due to non-tracked programs is to conduct surveys similar to those described in the Dougherty et al. 2011 paper "Moving Beyond Econometrics to Examine the Behavioral Changes Behind Impacts".

Specific Implementation for the HER Program

LBNL Observation 24 *LBNL recommends that PSE consider conducting survey research of customers to assess the possible impact of programs that are not tracked, such as upstream programs that cannot be traced to a specific household. For example, the analysis described in Dougherty et al. 2011[17] used surveys of individual households in order to determine the types of measures (e.g., appliances, CFLs, and weatherization) that households in a Massachusetts HER program installed.*

If the magnitude of the double counted savings cannot be estimated, then it is possible that this could cause a bias in estimates of savings; however, as described above, most of the double counted savings should be deducted from the other program. Therefore it is possible that the energy savings

[17] Doughterty, A., Dwelley, A., Henschel, R. and Hastings, R. *Moving Beyond Econometrics to Examine the Behavioral Changes behind Impacts.* IEPEC Conference Paper.

attributable to the CFL program are overestimated, and possible that energy savings attributable to the HER program are overestimated by a much smaller amount if at all.

7 Recommendations

Based on our analysis of the HER program, LBNL recommends the following going forward.

- In the future, LBNL recommends that a randomly allocated control and treatment group should be maintained in order to allow unbiased estimates of energy savings each year. In practice, this means that HERs cannot ever be mailed to *every* household. However, it is possible that the size of the control group could be reduced (50% of the study population is likely not needed). Analysis should be done to determine the smallest possible control group such that the estimates are likely to be statistically significant at 5%. If the control group size is reduced, then more people can be in the treatment group, and aggregate savings are likely to be higher[18].
- If the program is to be expanded to additional populations, LBNL recommends that a new study population is defined and a new control and treatment group randomly assigned from within the new study population. Again, an analysis should be done to determine the smallest possible control group so that as many households as possible can be in the treatment group.
- In future analysis, LBNL would recommend using either a difference-in-difference model defined in section 4.3.2 as Model 1 (which is the same as the method used to produce the "Difference-of-Differences" results in the last two columns of Table C-1 in KEMA's Evaluation), or a fixed effects regression with standard errors clustered at the household level that is not normalized for a "typical" year's weather, defined in section 4.3.2 as Model 2. Both models lead to unbiased estimates of energy savings with correctly calculated confidence intervals. Model 1 is a more simple analysis, while Model 2 may be slightly more precise in the sense that it may have slightly smaller confidence intervals.
- In the future LBNL recommends that KEMA or PSE continues to review program tracking databases to determine participation by customers in other PSE efficiency programs in order to calculate double counted savings for these tracked programs (using a method similar to that used in KEMA's Double Counting Memo).

[18] Two good references for determining the optimal control and treatment sample sizes are (1) section four in: Duflo, E., R. Glennerster, and M. Kremer. 2007. "Using randomization in development economics research: A toolkit." *Handbook of development economics* 4: 3895-3962. http://economics.mit.edu/files/806; and (2) section 4 Protocol 5 in: "Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols." EPRI, Palo Alto, CA: 2010. 1020855.

- LBNL recommends that PSE consider conducting survey research of customers to assess the possible impact of programs that are not tracked, such as upstream programs that cannot be traced to a specific household. For example, the analysis described in Dougherty et al. 2011[19] used surveys of individual households in order to determine the types of measures (e.g., appliances, CFLs, and weatherization) that households in a Massachusetts HER program installed.

[19] Doughterty, A., Dwelley, A., Henschel, R. and Hastings, R. *Moving Beyond Econometrics to Examine the Behavioral Changes behind Impacts.* IEPEC Conference Paper.