MINNESOTA DEPARTMENT OF
COMMERCE
DIVISION OF ENERGY RESOURCES

Audit of EnergyScoreCards Minnesota Impact Evaluation

A Pilot Program to Study Energy and Water Benchmarking in Multifamily Buildings

Conservation Applied Research & Development (CARD)
FINAL REPORT

Prepared for: Minnesota Department of Commerce,
Division of Energy Resources

Prepared by: Center for Energy and Environment

July 2015

cee
Center for Energy and Environment

Prepared by:

Martha Hewett, Center for Energy and Environment
Di Sui, Center for Energy and Environment
Russ Landry, Center for Energy and Environment
Carl Nelson, Center for Energy and Environment

Center for Energy and Environment
212 3rd Ave North, Suite 560
Minneapolis, MN 55401
612-335-5858
www.mncee.org
Project Contact: Carl Nelson, cnelson@mncee.org

Contract Number: 0519012

Prepared for Minnesota Department of Commerce, Division of Energy Resources
Mike Rothman, Commissioner, Department of Commerce
Bill Grant, Deputy Commissioner, Department of Commerce, Division of Energy Resources
Laura Silver, Project Manager
651-539-1873
Laura.Silver@state.mn.us

ACKNOWLEDGEMENTS

This project was supported in part (or in whole) by a grant from the Minnesota Department of Commerce, Division of Energy Resources, through the Conservation Applied Research and Development (CARD) program, which is funded by Minnesota ratepayers.

DISCLAIMER

This report does not necessarily represent the view(s), opinion(s), or position(s) of the Minnesota Department of Commerce (Commerce), its employees or the State of Minnesota (State). When applicable, the State will evaluate the results of this research for inclusion in Conservation Improvement Program (CIP) portfolios and communicate its recommendations in separate document(s).

Commerce, the State, its employees, contractors, subcontractors, project participants, the organizations listed herein, or any person on behalf of any of the organizations mentioned herein make no warranty, express or implied, with respect to the use of any information, apparatus, method, or process disclosed in this document. Furthermore, the aforementioned parties assume no liability for the information in this report with respect to the use of, or damages resulting from the use of, any information, apparatus, method, or process disclosed in this document; nor does any party represent that the use of this information will not infringe upon privately owned rights.

Table of Contents

1. Summary

This report documents the Center for Energy and Environment's (CEE's) audit of the EnergyScoreCards Minnesota (ESC Minnesota) impact evaluation. The impact evaluation itself was conducted by Bright Power, the implementer of the EnergyScoreCards Minnesota pilot project and the owner of the EnergyScoreCards' software. CEE's audit was conducted to provide third-party review of the evaluation, mitigating potential concerns about conflict of interest inherent in evaluations conducted by a program implementer.

Bright Power's evaluation included a utility bill analysis of treatment and control properties to assess the relative change in energy use between treatment and control groups (see separate report). The scope of CEE's audit subcontract included review of the following:

- Design and selection of the treatment and control groups;
- Data integrity and savings estimates for a small sample of cases;
- Methodology for computing the change in energy use of individual treatment and control properties; and
- Methodology for computing aggregate savings and uncertainty in aggregate savings.

In addition to the audit, CEE provided input and feedback to Bright Power at numerous stages in the study. CEE provided guidance on sample design and actually assigned the recruited portfolios to participant or control status, provided guidance on appropriate analysis techniques consistent with the sample design, provided a checklist of key items to address in the analysis and reporting, and provided extensive review of several rounds of statistical analysis and of two drafts of the impact evaluation. This audit was completed prior to Bright Power's completion of the final version of the impact evaluation, and is based on drafts of the report provided to CEE through April 21, 2015.

CEE's audit is intended to provide a third-party review of the data and methods used in the impact evaluation. It should not be construed as considering all potential areas of bias or concern, but represents CEE's best effort given budget and time constraints to review the most important aspects of the impact evaluation.

Overall, we found the approach used by Bright Power in the near final version of the EnergyScoreCards Minnesota impact evaluation to be reasonable and appropriate.

The sample design (a stratified cluster-randomized controlled trial) is appropriate for the multifamily market, which has quite diverse properties managed within portfolios. Portfolios were actually assigned to treatment or control status by CEE, to avoid the possibility of assignment bias, and were checked for balance of characteristics between the two groups. These factors ensure that the study has internal validity; that is, that the observed impacts were due to the pilot program itself, and not to other confounding factors. Given the opt-in sample design, the pilot project buildings are not a random sample from the population of multifamily buildings in Minnesota, and therefore the study does not have external validity; that is, the study results cannot be assumed to be an accurate estimate of the savings that would be achieved in delivering a full scale program to a random sample of Minnesota multi-unit housing. The results can be used, however, to make more informed judgments as to the level of savings that might occur in such a program.

We found no indication of systematic bias in the calculation of savings for the individual buildings. We independently replicated the indices used by Bright Power to calculate individual savings estimates for three buildings in the pilot, using data independent of that collected by Bright Power, and using Bright Power's methodology for calculating building energy usage.

CEE provided considerable guidance to Bright Power in designing, conducting and reporting the statistical analysis of aggregate savings. The final analysis appropriately states the null hypothesis and uses standard rejection criteria, uses a full year of pre-program data as a baseline, minimizes sample loss, uses an accurately prepared data set, properly checks equivalency of the treatment and control properties remaining in the sample at the time of analysis, properly specifies analysis models suitable for a cluster-randomized controlled trial, avoids double-counting of savings achieved due to increased uptake of utility rebates by portfolios in the treatment group, appropriately calculates the chosen cost-effectiveness metric, and fully and clearly reports results. While there are a few aspects of the analysis that are not fully optimized, our judgment is that they would be unlikely to materially change the conclusions drawn. Based on our extensive review of the statistical and other analysis and of the drafts of the evaluation report, we believe it to be an accurate reflection of true pilot project outcomes.

Retrospective analysis of statistical power suggests that the pilot project sample and sample design may have been inadequate to detect statistically significant savings of the magnitude suggested by regression results for the pilot group as a whole. Thus, it may be more appropriate to conclude that the sample was not large enough to detect savings realized by the entire pilot group than to conclude that there were no overall project savings.

2. Design and Selection of the Treatment and Control Groups

The design and selection of treatment and control groups has critical implications for the validity and statistical power of a pilot project:

- The *internal validity* of study results – the extent to which it can be determined that the observed pilot project energy and water savings were actually caused by the pilot project, rather than some other factor(s) – depends on the experimental design and on the absence of experimenter bias.

- The *external validity* of study results – the extent to which the observed pilot project results can be generalized to other groups of buildings – likewise depends on study design.

- The *statistical power* of the hypothesis tests used to analyze study outcomes – their ability to detect savings -- depends on both the size of the sample included in the pilot project and the sample design.

Each of these issues is discussed in turn below.

2.1. Internal Validity

For study results to be internally valid, the experimental design must ensure that the observed impacts were not due to some other confounding factor, were not due to selection bias (pre-existing differences between treatment and control buildings) and were not due to cross-contamination between the treatment and control groups. Bright Power chose a randomized controlled trial design, in which the subjects are randomly assigned to participant or control status, in order to minimize the potential for confounding effects and selection bias. Such designs are commonly considered to be the "gold standard" in experimental design, and in particular have been recommended for design of behavior-based programs.[1] In order to minimize the possibility of experimenter bias in the assignment of buildings to treatment or control status, it was determined that CEE, rather than Bright Power, would carry out the assignment.

In their proposal Bright Power had anticipated simple random assignment. However, in early project planning,[2] CEE pointed out that there was significant potential for cross-contamination if buildings in a given owner's or management company's portfolio were divided between treatment and control status. There would be no way to ensure that conservation actions an owner (or manager) learned about from the account managers, or decided to undertake in his "participant" buildings based on the EnergyScoreCards information, would not also be applied to his "control" buildings. Since the fundamental premise of the EnergyScoreCards service is that the owner (manager) will take action to

[1] State and Local Energy Efficiency Action Network (SEE Action Network). 2012. *Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations*. Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman, Lawrence Berkeley National Laboratory. http://behavioranalytics.lbl.gov.

[2] October 2011.

change building operation or invest in capital equipment based on feedback and other information obtained through the service, the owner (manager) must be the unit of randomization. All buildings in a given owner’s (manager’s) portfolio must be assigned to either treatment or control status, rather than a mixture of statuses. Analytically, this means that each participating organization is treated as a “cluster” and that the study buildings are “cluster-randomized” or “group-randomized.” In addition, the project team informed CEE that nineteen of the organizations had close relationships with some other organizations such that cross-contamination might occur between organizations. To address this, these organizations were grouped into seven “super-clusters,” and these super-clusters were assigned together to either treatment or control status.

In general there are two options to improve precision in cluster-randomized experiments – blocking, in which clusters are first subdivided on key characteristics and then assigned randomly to participant or control status within blocks, and covariance adjustment, in which adjustments are made during analysis by treating the outcome as a linear function of treatment group and other explanatory covariate(s) (also known as control variables). Blocking causes a greater loss in degrees of freedom and can actually produce worse precision than strictly random assignment of clusters under some circumstances. So, in general, covariance adjustment is preferable. However it can be useful to do some blocking in order to improve the face validity of the sample and to avoid having completely disparate treatment and control samples when the clusters are highly heterogeneous, as is the case here with the portfolios. Owners (managers) have a much greater motivation to reduce energy use on the accounts for which the owner pays the bills. For that reason, it was decided to block applicant portfolios based on the predominant payment code of buildings in the portfolio (payment code is a coding created by Bright Power that describes the energy end uses for which owners or tenants pay the bills). This resulted in three blocks, or strata. The largest block is primarily (86%) payment code (T)TOO (tenant pays for apartment electricity and cooling, owner pays for heating and hot water), the second largest is primarily (74%) payment code (O)OOO (owner pays for everything), and the third, smallest, block is 50% (T)TTO and 9% (T)TTT (in both of these payment types the tenant pays for apartment electricity, cooling and heating). Clusters (or super-clusters, where those existed) were randomly assigned to treatment or control status within these blocks (strata).

After a trial randomization, it is good practice to compare the resulting treatment and control groups to ascertain whether they are reasonably well balanced with respect to factors considered likely to be related to the primary experimental outcome (in this case, change in energy use). CEE conducted this “equivalency check” both on factors we thought might be related to the experimental outcome and on factors we thought might be important to the face validity of the study (acceptability of the treatment/control assignment to key stakeholders). The trial randomization was reviewed with the project partners and deemed to be acceptable based on the observed degree of balance in these factors.

The resulting design is considered a stratified cluster-randomized controlled trial, a design which is widely used. The stratification and clustering must be taken into account in the analysis phase in order to obtain correct estimates of the standard error of savings and the statistical significance of those savings. This issue is discussed further in Section 5 of this audit.

Further details on the assignment of candidate buildings to treatment or control status and on the results of the equivalency checks are given in Appendix B of Bright Power's "EnergyScoreCards Minnesota Impact Evaluation." This appendix was prepared as a working document by CEE at the time of participant and control group assignment.

The sample design was not strictly random in one respect, attributable to practical considerations. Two portfolios were forced into the participant group, since they already had experience with EnergyScoreCards and had some of the properties submitted as candidates for the pilot in the ESC tool already. Bright Power did not consider it "politically" feasible to exclude these portfolios from participation in the pilot project, and in addition was interested in including the buildings to maximize sample size. One of the two portfolios had 15 properties in the study and was part of a super-cluster that also had 12 other properties in the study. The other portfolio had 5 properties in the study. It is impossible to know with the information available to us whether these portfolios would be more likely to experience savings (e.g., because they were already familiar with ESC), or less likely to experience savings (e.g., because the major impact of benchmarking may have occurred prior to the study for their properties). Since Bright Power only found measurable savings for buildings with payment code (O)OOO, it is useful to know how many of the properties in these portfolios had this payment code. In the two portfolios themselves, there were four properties with this payment code. In the other portfolios that were part of the same super-cluster as one of these portfolios, there were an additional five properties with this payment code. (It is possible that fewer than nine of these (O)OOO properties had sufficient valid data to remain in the final analysis sample.) One way that the impact of this practical non-random assignment could be assessed would be to drop the 32 properties involved from the sample and repeat the key analyses to see if the results are substantially different. To date that has not been done.

The cluster-randomization, the blocking and resulting balance on key variables, and the assignment to treatment or control status by CEE, a third party with no investment in the outcome of the pilot project, are believed to have resulted in a study with strong internal validity; that is, the impacts, if determined through proper analysis, are believed to be attributable to the pilot project itself, and not to confounding factors, selection bias, or other causes. One weak point in the internal validity is the forced assignment of two portfolios to the treatment group. It is unclear whether this would have tended to increase or decrease the observed savings. This was not examined in Bright Power's analysis.

2.2. External Validity

Pilot project results would be externally valid if they could be reliably generalized to other situations, for example, another group of buildings in a full scale program. In general, results cannot be reliably extrapolated to other groups unless an experimental intervention is delivered to a random sample from a defined population, and the other groups in question are also drawn at random from that population. Obviously this implies that the results of the present pilot project cannot reliably be extended to buildings in other states (where the climate, rental property size, age and condition, rental property owner/manager characteristics and attitudes, utility prices, utility rebate programs or other factors may differ from those in Minnesota) or to other time periods (for example, when the price of energy, the rate of change in those prices, the vacancy rate or other factors affecting the rental housing industry may differ).

Moreover, the candidate buildings for the pilot project were not a random sample of multifamily buildings in Minnesota, and thus the results cannot reliably be extrapolated to other multifamily buildings in the state. Specifically, the portfolios included in the project were a convenience sample that opted in to the study. Outreach was conducted through a public website, a series of open webinars, e-mail blasts sent out by the Minnesota Housing Finance Agency (MHFA) and the Minnesota Multi Housing Association (MHA), and a phone and e-mail campaign contacting eligible owners based on lists from MHFA. Thus it is likely that owners/management companies not on either MHFA or MHA lists did not hear about the program. After an owner or management company decided to opt in, their buildings had to meet certain eligibility criteria in order to be accepted into the pilot project. Specifically, they had to be located in the service territories of Xcel Energy, CenterPoint Energy, and/or Rochester, Austin, Owatonna, St. Cloud and Mankato utilities (to facilitate collection of utility data), have 10 or more units, not be townhomes, condos or cooperatives, not be under a performance contract, and have at least ten months of existing "operating history"[3] as of the pilot launch in early 2012. In addition, the owner (manager) had to complete the required property survey and authorize ESC Minnesota to collect owner-paid utility data for the duration of the pilot.

Initially, owners/managers were allowed to enroll only a small number properties from their portfolio. This rule was established because the precision of savings estimates in a cluster-randomized trial is influenced more heavily by the number of clusters than the total number of buildings, and it was therefore recommended by CEE that the project team enroll more portfolios and limit the number of buildings per portfolio. Later this criterion was relaxed because of difficulties recruiting the promised size sample within the time available.

Clearly, given the opt-in recruitment, the specifics of the marketing and the eligibility criteria, the resulting pilot project sample is not a random sample of multifamily buildings in Minnesota. There is therefore no assurance that buildings enrolled in some future program would see results similar to those in this program. Therefore, projections about the outcomes of such potential programs should be viewed with caution. Savings could well be lower or higher than observed in the pilot. Pilot project results may reasonably be used to inform a judgement as to whether to offer such a program, but actual results would need to be determined through an experimental design with a control group and analysis similar to that used for the pilot.

Likewise, the current pilot project provides no information with which to estimate the *long term* impacts of a full scale EnergyScoreCards program. The pilot project as a whole did not show significant savings in either year, though one subgroup, those with payment code (O)OOO, showed significant savings in the second year. In their analysis of cost-effectiveness, Bright Power presents one scenario in which they extrapolate results for a 10 year program and assume that savings begin to occur in the second year and

[3] This presumably includes a minimum 10 month period during which the present owner/manager was responsible for the utility bills and could therefore grant a release for those data. A building which had recently changed hands, then, would not qualify.

continue at the same level for years 3 through 10. While this is perhaps reasonable as a hypothetical scenario, the pilot project itself does not provide results that can be extrapolated to future years without continuing the project, and its participant and control groups, for a longer period.

Thus the pilot project lacks external validity, and the results can only be extrapolated to other situations in a qualitative way, to inform decisions that must be made where directly applicable results are not available.

2.3. Statistical Power

The statistical power of a pilot project of this type is its ability to detect savings (i.e., to reject the null hypothesis that savings are zero) when in fact there are savings. While the analysis at the conclusion of a project ensures that savings are not claimed if they are not statistically significant, an equally important concern is that savings be recognized when they do occur. Ideally, this concern should be addressed during the planning phase by ensuring that the sample size and design have enough power to detect an effect of the size the researcher expects to see.

Statistical power depends on both the size of the sample and the sample design. For the present project, the total sample size was established by budget constraints and, presumably, by Bright Power's qualitative sense of the number of buildings that might be required to allow savings to be detected, rather than through a formal power analysis. The decision to use a cluster-randomized design (in order to minimize the risk of cross-contamination, as discussed above) has important implications for statistical power. In a cluster-randomized trial, the number of clusters has a greater impact on power than does the total sample size. For this reason, CEE originally recommended that the maximum number of buildings per owner included in the pilot be strictly limited, with the 500 buildings to which Bright Power had committed in their proposal representing at least 100 portfolios and preferably more. Due to difficulties in meeting the 500 building goal and the fact that many owners (managers) submitted large numbers of candidate properties, the maximum number of buildings accepted per portfolio was gradually increased, first to 12, then to 15. In the final sample, the number of properties per portfolio ranged from 1 to 24, with a mean of 6.1 and a median of 4. The final sample had a total of 564 properties in 93 portfolios. The analysis sample was smaller than this due to lack of sufficient utility data for some buildings in some years.

Statistical power also depends on factors such as the average savings and the variability in savings across buildings, which cannot be controlled by the experimenter. It depends, too, on whether the experimenter is able to identify and measure covariates that explain some of the variability in savings and can be used to increase the precision of the estimated program impact. At the start of this project, no information was available with which to estimate the likely mean or variance of savings or the predictive power of available covariates. At the conclusion of the project, however, estimates of such parameters were available, and these were used by Bright Power, with guidance from CEE, to estimate the sample size that would have been required to allow savings to be detected. This work is discussed further in Section 5. The information obtained from this power analysis will be valuable in planning future projects using the EnergyScoreCards service, and also provides valuable context regarding the non-significant savings in this project.

3. Data Integrity and Savings Estimates for a Small Sample of Cases

3.1 The EnergyScoreCards Methodology for Calculating Savings Estimates

Bright Power used the indices calculated by the EnergyScoreCards software for calculating pre and post savings in the treatment and control groups (see impact evaluation for full description). Thus, the calculation of the indices is the basis for calculating the savings estimates. This section describes the methodology used by Bright Power to calculate the indices, which was replicated by CEE as described below. The EnergyScoreCards calculations are performed within the EnergyScoreCards software, while CEE replicated the calculations in Microsoft Excel.

3.1.1 Data Preparation

The EnergyScoreCards indices were calculated using monthly metered data for each building and daily weather data. The monthly metered data for each building was pulled from either a utility company website via a “data scraper” function within EnergyScoreCards, or from a direct utility feed, as was the case in Minnesota with Xcel Energy. The daily weather data came from NOAA and was used to calculate Heating Degree Days (HDD) and Cooling Degree Days (CDD) with a balance point of 65°F. The HDD and CDD calculations for each billing period were based on the dates of the reported meter readings. For each calculation the meter reading end date was included, but the start date was not.

3.1.2 Calculation Methodology

Regression Analysis for Each Fuel Type

A multi-variable ordinary least squares (OLS) regression was used for each account to determine the heating, cooling and non-weather dependent components of energy use on that account for each calendar year in the study. Billing months that spanned two calendar years were included in the second of the two years. The independent variables used in the linear regression model were HDD, CDD, and days in the billing cycle, with no zeroth order term (constant). A number of buildings had more than one account for either or both gas and electricity. For each fuel type (gas and electricity), the metered data was broken up into different regression components, represented by corresponding indices.

The regression processes were adjusted for the different fuel types. The regression components, processes and equations for each fuel type are listed below:

<u>Gas</u>

1. Regression Components:
   Heating and non-weather dependent components.

2. Regression Process:
   The fossil fuel baseload index was estimated for each year by averaging the daily gas usage of the summer months that have zero or minimal HDD (< 1 HDD per day). Base loads were then

calculated and extracted from each month before conducting an OLS regression with HDD only to determine the heating index.

3. Regression Equations:

$$\text{Fossil Fuel Baseload Index} = \frac{\text{Sum of Usage of Low HDD Months}}{\text{Sum of Days of Low HDD Months}}$$

$$\text{Adjusted Usage(Bill)} = \text{Original Usage(Bill)} - \text{Days(Bill)} \times \text{Fossil Fuel Baseload Index}$$

$$\text{Usage(Bill)} = \text{HDD(Bill)} \times \text{Heating Index}$$

<u>Electric</u>

1. Regression Components:
   Cooling and non-weather dependent components, or heating, cooling and non-weather dependent components.

2. Regression Process:
   A normal OLS regression was conducted with and without heating, and the model that had the best fit without having any negative coefficients was selected.[4]

3. Regression Equation:

$$\text{Usage(Bill)} = \text{CDD(Bill)} \times \text{Cooling Index} + \text{Days(Bill)} \times \text{Electric Baseload Index}$$

Or

$$\text{Usage(Bill)} = \text{HDD(Bill)} \times \text{Heating Index} + \text{CDD(Bill)} \times \text{Cooling Index} + \text{Days(Bill)} \times \text{Electric Baseload Index}$$

For both fuel types, individual monthly utility bills were assessed using the regression results to determine whether they should be included in the evaluation. Any bill that was more than a specified amount greater than the regression standard error from the model was eliminated and the regression for that meter was then recalculated without that month of data.

Unit Conversion and Property Summary

Regression analysis for each account's data was performed separately for each account. Once the regression processes were complete, the total energy usage was added together for each property. The regression results were then converted into standard units to calculate the final energy usage indices. To convert the heating and cooling indices into a standard unit, the calculations used the building areas that were paid for by the building owner (this could be just the common area or the entire building).

[4] ESC reported that electric usage is often best modeled with heating component even when electricity is not reported as a heat source.

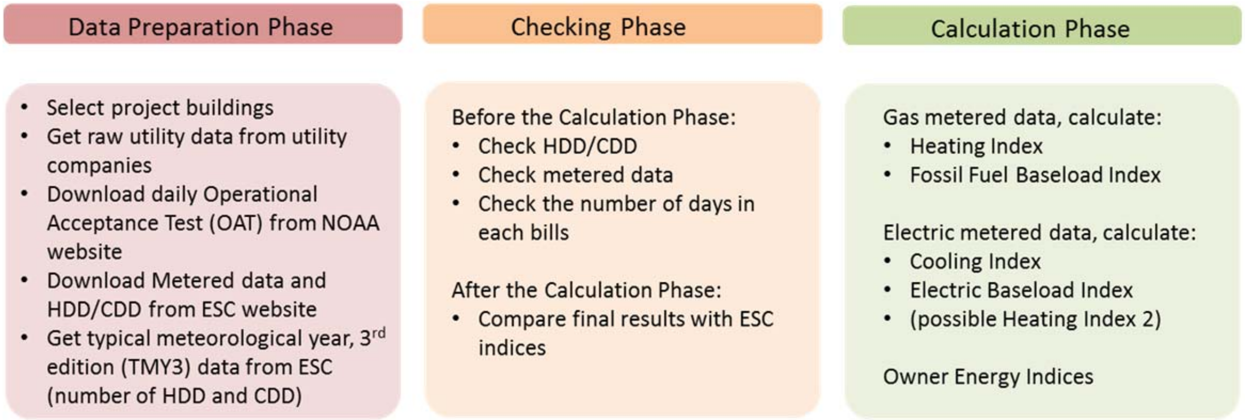
Owner Energy Use Index (EUI)

Weather normalized Owner EUIs were generated by combining the above indices with HDD and CDD from TMY3 data representing a typical weather year. EUIs were generated for each year for each property.

3.2 Verification Methodology

The verification process consists of three phases, as shown in Figure 1.

Figure 1. CEE Verification Process



Data Preparation Phase

- Select project buildings
- Get raw utility data from utility companies
- Download daily Operational Acceptance Test (OAT) from NOAA website
- Download Metered data and HDD/CDD from ESC website
- Get typical meteorological year, 3rd edition (TMY3) data from ESC (number of HDD and CDD)

Checking Phase

Before the Calculation Phase:

- Check HDD/CDD
- Check metered data
- Check the number of days in each bills

After the Calculation Phase:

- Compare final results with ESC indices

Calculation Phase

Gas metered data, calculate:

- Heating Index
- Fossil Fuel Baseload Index

Electric metered data, calculate:

- Cooling Index
- Electric Baseload Index
- (possible Heating Index 2)

Owner Energy Indices

For the verification process, CEE followed Bright Power's steps in order to duplicate their calculations and assess the data and computations. These calculations were replicated for three buildings. We note that this is a relatively small sample size compared to the number in the overall pilot, and may not be a large enough sample size to conclusively eliminate the possibility of bias in the results. However, we believe that our deep dive into a few randomly-selected cases can provide a reasonable assessment of whether there is likely to be any systemic bias in the approach used. Detailed information of each of the steps is described as below.

3.2.1 Data Preparation Phase

Three buildings were selected for testing ("Project Buildings 1, 2 and 3"). These buildings were chosen randomly from a pool of buildings that were within the top or bottom 5 percent of estimated annual energy savings from the pre to post periods. In duplicating the work of Bright Power, daily outdoor air temperature from NOAA and metered data from the EnergyScoreCards website were downloaded for each of the three project buildings. To verify the data integrity, the HDD and CDD numbers used by Bright Power were obtained from the EnergyScoreCards online software, and raw utility data for the three buildings were obtained directly from the utility companies to be used for comparisons. TMY3 data (which included HDD and CDD values) were obtained directly from Bright Power and are listed in Table 1.

Table 1. TMY3 HDD and CDD Values

Location	Typical HDD	Typical CDD
Minneapolis	7721.45	756.44
Duluth	9565.46	162.02
Sioux Falls	7612.41	734.35

3.2.2 Checking Phase

A series of checks were performed on the data collected in the first phase before calculations were completed. Table 2 lists the data that was checked in this phase, along with the data sources for calculation and comparison.

Table 2. Checked Items for Data Integrity

Items	Data Used for CEE Independent Calculation	Data Compared to
HDD/CDD	Calculated from NOAA mean daily temperature	Recorded on ESC website (in Post-Checking Phase)
Metered Data	Downloaded from ESC website	Raw utility data from utility company
# Days Billed	Same billing period as shown on ESC website Each period includes the end date, not the start date	-
Indices Results	-	Compare the CEE calculated indices with ESC results

3.2.3 Calculation Phase

Calculations of the six indices were conducted using Bright Power's methodology. The regression processes were carried out using the multiple regression function in Microsoft Excel.

3.3 Verification Results

3.3.1 HDD and CDD

The following equation was used to calculate the difference between the HDD and the CDD from NOAA and from the ESC website:

$$\text{Difference } = \frac{\text{NOAA} - \text{ESC}}{\text{ESC}} \qquad (\%)$$

Table 3 shows the results of the calculation for each project building. Except for the HDD values for the gas meter calculations of Project Building 2, the checking results showed similar variation between the buildings during the same year and with same meter type. Four checks had a variation larger than 1%, and are marked in red in Table 3. Compared to other indices, the cooling indices are much smaller and have less impact on the total EUI and thus the three differences greater than 1% that occurred in 2012 did not have a large impact on the calculated EUI values. The impact of the variations that are less than 1 percent (the HDD in 2012 for the gas meter of Project Building 2) will be explained in later comparisons.

Table 3. HDD and CDD Checking Results

Heating Degree Days (Base 65)						
Account Type	Gas			Electric		
Year	2012	2013	2014	2012	2013	2014
Project Building 1	-0.56%	-0.38%	0	-0.56%	-0.30%	0
Project Building 2	-2.50%	-0.33%	-0.34%	-0.51%	--	0
Project Building 3	-0.54%	-0.29%	0	-0.59%	-0.30%	0
Cooling Degree Days (Base 65)						
Project Building 1	-	-	-	1.95%	0.78%	0
Project Building 2	-	-	-	1.95%	0.78%	0
Project Building 3	-	-	-	1.95%	0.78%	0

3.3.2 Metered Data

The metered data downloaded from the ESC website were checked against the raw utility data. The results are shown in Table 4.

Table 4. Metered Data Checking Results

Building Name	**Item Checked**	**Checking result**
Project Building 1	Gas usage(02/08/12-12/28/14) Electric usage(01/07/12-12/27/14)	Match
Project Building 2	Electric usage(02/29/12-12/31/14)	Missing one month in ESC
Project Building 3	Electric usage(12/15/12-12/14/14) Gas usage(01/14/12-12/05/14)	Match

All metered data that was recorded on the EnergyScoreCards website for the three project buildings were checked, except the gas usage of Project Building 2. As shown in Table 4, one monthly electric bill was missing from the ESC system for Project Building 2. The regression results with and without this missing month were very similar (Table 5).

Table 5. EUI Results With and Without the Missing Month Data

	ESC	**CEE w/o missing month**	**CEE w/ missing month**
EUI	37.49	37.36	37.61
diff%	-	-0.36%	0.32%

*diff% = (CEE-ESC)/ESC

3.2.3 Bad Bills

The monthly metered data were evaluated by comparing the residuals to EnergyScoreCards' threshold for screening out the regression standard errors.[5] Two bad bills were found in the check:

[5] See Section 5.3 for a complete description of the process Bright Power uses to screen out potentially incorrect data from the analysis.

Table 6. Bad Bills Checking Results

Building Name	Meter Type	Billing Year
Project Building 1	1 month of gas bill	2013
Project Building 2	1 month of electric bill	2014

3.3.3 Index Results

The comparisons of the indices calculated by CEE with those calculated by Bright Power are shown in Table 7.

Table 7. Index Calculation Results

	Owner Energy Index (kBTU/ft^2/yr):								
	2012			2013			2014		
	ESC	CEE	**diff%**	ESC	CEE	**diff%**	ESC	CEE	**diff%**
Project Building 1	82.29	82.44	**0.18%**	84.69	84.61	**-0.09%**	52.65	52.64	**-0.02%**
Project Building 2	32.34	32.58	**0.75%**	34.12	34.15	**0.09%**	37.49	37.55	**0.15%**
Project Building 3	183.77	184.21	**0.24%**	165.85	166.04	**0.11%**	157.98	157.95	**-0.02%**
	Heating Index (BTU/HDD/ ft^2)								
	2012			2013			2014		
	ESC	CEE	**diff%**	ESC	CEE	**diff%**	ESC	CEE	**diff%**
Project Building 1	7.58	7.60	**0.30%**	7.92	7.94	**0.33%**	3.92	3.92	**-0.03%**
Project Building 2	2.55	2.58	**1.28%**	2.63	2.63	**0.17%**	2.59	2.60	**0.26%**
Project Building 3	18.81	18.87	**0.32%**	16.50	16.53	**0.18%**	15.25	15.24	**-0.02%**
	Electric Baseload Index (kWh/unit/yr)								
	2012			2013			2014		
	ESC	CEE	**diff%**	ESC	CEE	**diff%**	ESC	CEE	**diff%**
Project Building 1	4230.84	4235.91	**0.12%**	4019.25	4002.77	**-0.41%**	3938.10	3938.10	**0.00%**
Project Building 2	2131.98	2116.09	**-0.75%**	2186.24	2185.44	**-0.04%**	2028.71	2059.00	**0.00%**
Project Building 3	2187.80	2135.66	**0.15%**	2441.17	2182.09	**-0.17%**	2379.65	2028.74	**0.00%**
	Fossil Fuel Baseload Index (mmBTU/bedroom/yr)								
	2012			2013			2014		
	ESC	CEE	**diff%**	ESC	CEE	**diff%**	ESC	CEE	**diff%**
Project Building 1	6.50	6.50	**-0.02%**	7.29	7.29	**-0.02%**	6.14	6.14	**-0.02%**
Project Building 2	6.55	6.55	**-0.02%**	7.44	7.43	**-0.02%**	10.85	10.85	**-0.02%**
Project Building 3	6.76	6.76	**-0.02%**	6.54	6.54	**-0.02%**	7.15	7.15	**-0.02%**

*diff% = (CEE-ESC)/ESC

A substantial difference was found, as highlighted in Table 7, for the heating index for Building 2. A possible cause for this variances is the difference between HDD values used in the CEE calculation and those used by Bright Power, as mentioned in the checking results for the HDD values in table 3.

An adjusted index calculation was conducted using the ESC HDD and whole building area for cooling indices. The results are shown in Table 8. This reduced all of the variations to less than 1 percent.

Table 8. Comparisons of Adjusted Indices

	Owner Energy Index (kBTU/ft^2/yr):								
	2012			2013			2014		
	ESC	CEE	**diff%**	ESC	CEE	**diff%**	ESC	CEE	**diff%**
Original Project Building 2	32.34	32.58	**0.75%**	34.12	34.15	**0.09%**	37.49	37.55	**0.15%**
Corrected Project Building 2	32.34	32.33	**-0.04%**	34.12	34.09	**0.09%**	37.49	37.55	**0.15%**
	Heating Index (BTU/HDD/ft^2)								
	2012			2013			2014		
	ESC	CEE	**diff%**	ESC	CEE	**diff%**	ESC	CEE	**diff%**
Original Project Building 2	2.55	2.58	**1.28%**	2.63	2.63	**0.17%**	2.59	2.60	**0.26%**
Corrected Project Building 2	2.55	2.55	**-0.02%**	2.63	2.63	**0.17%**	2.59	2.60	**0.26%**

While the variations in a given year were less than 1 percent, Bright Power's overall analysis of the treatment and control groups considered the energy use differences from year to year to estimate the savings (specifically, the differences in energy use from 2012 to 2014). This would tend to magnify the impact of the small differences found here. As shown on Table 9, the savings from 2012 to 2014 for the three test buildings varies from about negative 16 percent to positive 36 percent. The difference in savings as calculated by EnergyScoreCards versus CEE shows an OEI difference of 0.16 to 0.47, or 0.5% to 3.5%. Note that as CEE chose buildings with a high change from 2012 to 2014, these differences may be higher than the average.

Table 9. Comparison of EnergyScoreCards-calculated and CEE-calculated 2012-2014 Owner Energy Index Savings

	2012 OEI		**2014 OEI**		**2012 to 2014 savings**		
	ESC	*CEE*	*ESC*	*CEE*	*ESC*	*CEE*	*Difference*
Project Building 1	82.29	82.44	52.65	52.64	29.64 (36.0%)	29.8 (36.1%)	-0.16 (-0.5%)
Project Building 2	32.34	32.58	37.49	37.55	-5.15 (-15.9%)	-4.97 (-15.3%)	-0.18 (3.5%)
Project Building 3	183.77	184.21	157.98	157.95	25.79 (14.0%)	26.26 (14.3%)	-0.47 (-1.8%)

This analysis indicates the possibility for a positive bias in the Bright Power calculation of savings; that is, in the three buildings analyzed here, the CEE calculations of savings were all lower than the Bright Power calculations of savings. However, the magnitude is not great, and the likely impact on the overall results is likely small.

3.3.4 Issues Observed

There is a consistent difference in HDD and CDD between the data downloaded by CEE from NOAA and the EnergyScoreCards website, especially for the CDD value in 2012, which has a deviation of 1.95%. In the savings analysis between the treatment and control groups, this should not introduce significant bias.

4. Evaluation of Methodology for Individual Building Calculations

An assessment of Bright Power's methodology was performed for the following two issues that were initially considered to have a potential impact on the savings analysis for the individual buildings:

1. The use of 65°F as a balance point temperature for the calculation of both HDD and CDD for all buildings.

2. The adjusted regression process used by EnergyScoreCards for gas data. The EnergyScoreCards method of evaluating and subtracting baseload usage from the metered data prior to performing the regression is a non-standard statistical analysis approach that could have a significant impact on indices and total estimated energy use.

4.1 Fixed Balance Point

The use of a fixed 65°F balance point temperature is a possible cause of error and bias in the evaluation. This is because a 65°F balance point does not represent all buildings and in these instances, year to year variations in weather could lead to apparent differences in both the indices and EUI. However, these errors are only significant for a limited number of buildings, and the comparison to the control group should make their impact very small. Recent analysis by Energy Center of Wisconsin[6] suggests that most multi-family buildings in Minnesota's climate zone are fairly well represented by the 65°F balance point. Their floating balance point calculation on 86 multi-family buildings in Minnesota with master-metered gas data showed an average balance point of 65°F. Thus, we consider the use of 65°F for balance point in the Bright Power's analysis as reasonable and unlikely to have a significant impact on the comparison between treatment and control groups.

4.2 Adjusted Regression Process for Heating Index Calculation with Metered Gas Data

Since under most situations, the heating energy use covers the major part of the total building energy use, it is essential to verify that the method of gas baseload use and heating index calculation is reasonable. Considering that gas is the major owner-paid fuel for most of the buildings, and the regression process for gas metered data that Bright Power used is uncommon, a conventional OLS regression with heating and fossil fuel base load was conducted for each gas for all three project buildings. Table 10 shows the parameters that were compared between the two models.

[6] Personal communication with Scott Pigg, 3/5/15.

Table 10. Comparison Factors between Two Models

Name	**Multivariable Regression**	**ESC Regression Approach**
Heating Model Error Indicator	Regression standard error	Regression standard error
Baseload Model Error Indicator	$\frac{Baseload\ coefficient\ standard\ error}{Baseload\ coefficient}\%$	$\frac{\text{Standard deviation of daily baseload} *}{\text{Average daily } baseload *}\%$

* Calculated based on the variations in daily base load between multiple months of summer data. Summer months are defined by ESC as those with HDD less than 1HDD/day.

The results of the comparison between two regression models are shown in Table 11. Although the regression standard error of the ESC adjusted model is 4.5% higher than the standard model, the standard deviation of the base load index calculation is on average 15.9% lower. The significant gain in accuracy of the base usage estimate is considered to more than offset the modest decrease in accuracy of the heating usage. Moreover, the ESC adjusted approach was observed to provide much more consistency in year to year comparisons of base usage for the same building. Thus, we consider the ESC adjusted regression model approach used by ESC to be reasonable and do not expect that it introduced a significant bias in the impact evaluation (especially when the use of a control group is considered).

Table 11. Standard Error Comparisons between Two Models

	Heating Model Error Indicator			**Base Model Error Indicator**		
Regression Model	Control	Adjusted	**Diff % ***	Control	Adjusted	**Diff % ****
Project Building 1	938	1047	**11.6%**	30%	9%	**-20.9%**
Project Building 2	273	286	**4.8%**	20%	18%	**-2.1%**
Project Building 3	374	363	**-3.0%**	31 %	8%	**-24.9%**
Average	528	565	**4.5%**	28%	12%	**-15.9%**

* Diff% was calculated as $(\text{Adjust} - \text{Control})/\text{Control}$

** Diff% was calculated as $\text{Adjust} - \text{Control}$

5. Methodology for Computing and Reporting Aggregate Savings and Uncertainty

Drawing accurate conclusions about the impact of a pilot project requires that the appropriate analysis methods be used. This involves many considerations, including (but not limited to):

- Formulating the null hypothesis and establishing criteria for rejecting it,
- Establishing the length and date ranges of the baseline and treatment periods,
- Deciding which accounts to include in the analysis,
- Accurately preparing the data for analysis,
- Checking equivalency of the treatment and control groups remaining in the analysis,
- Properly specifying the analysis models to be used,
- Avoiding double-counting of savings achieved due to increased use of utility rebate programs by portfolios in the treatment group,
- Assessing the persistence of savings, if possible,
- Calculating the cost-effectiveness of savings, and
- Fully and clearly reporting results.

Each of these issues is discussed in turn below.

At the time of this project, Bright Power had not had occasion to conduct an evaluation similar to that required here.[7] Therefore, CEE provided extensive guidance and review to assist them in obtaining reliable results and in reporting sufficient details in the evaluation report to allow readers to understand and assess the project results. This included:

- Reviewing Bright Power's first year evaluation,[8]
- Discussing key analysis requirements and recommended statistical software with Bright Power's vice president in charge of analytics,
- Providing references on cluster-randomized trials, statistical software options, and behavioral program evaluation guidance to Bright Power's vice president in charge of analytics,
- Recommending appropriate analysis methods,
- Providing a checklist of the items and issues that CEE would be looking for in the final impact evaluation,
- Reviewing and providing feedback on at least two rounds of statistical analysis, and
- Reviewing and providing feedback on two drafts of the impact evaluation report.

[7] Their only somewhat similar study had been an analysis of savings from two multifamily retrofit programs: J. Braman, S. Kolberg, and J. Perlman, 2014. Energy and Water Savings in Multifamily Retrofits.

[8] J. Braman, C. Kling, 2014. EnergyScoreCards Minnesota First Year Evaluation Report.

This guidance had a substantial impact on the conclusions drawn from the impact evaluation and on the presentation of evaluation results, as described below.

5.1. Null Hypothesis and Significance Criteria

In experiments of this type, an important standard practice is to establish a null hypothesis -- that is, a formal statement of the question that is being tested – and to state before conducting any analyses the criteria that will be used to reject the null hypothesis and accept the alternative hypothesis. In this case, the null hypothesis is that the pilot project did not save any energy, and the alternative hypothesis is that it did. CEE encouraged Bright Power to be clear in stating the null hypothesis, and throughout the report, that the intervention being tested is not just the EnergyScoreCards benchmarking tool, but the tool and all of the ancillary support services, including, but not limited to, the support provided by the account managers. CEE further encouraged Bright Power to use the generally accepted statistical significance level of 5% to reject the null hypothesis. These recommendations are reflected in the statement of the main hypothesis in the report, the definition and use of the term "EnergyScoreCards service" throughout the report, and the criteria used to identify and report on significant results.

5.2. Length and Date Ranges of Baseline and Treatment Periods

Because energy use is strongly related to weather, it is typically recommended that projects of this type include at least one year of baseline (pre-project) data and at least one year of treatment (project) data.[9] Bright Power did obtain a nominal year of pre data for all properties, and included two nominal years of project data. The inclusion of two years of treatment made sense because with investment properties, owners (managers) often cannot make major investments in energy efficiency immediately, but rather need to go through a new budget cycle in order to budget for and then implement these measures. In addition, two project years of data allowed the evaluation to provide at least some insight into how savings change over time.

Initially Bright Power had some difficulty finding a way to extract the number of months of complete utility data from their EnergyScoreCards software, but, following up on CEE requests, they were able to find a way to extract this information and planned to include it in the final report.

[9] See, for example, SEE Action Network, op. cit.

Table 12 shows a breakdown of the percent of properties with various numbers of months of complete data in 2012 (the nominal pre year) and 2013 and 2014 (the nominal post years), based on data provided by Bright Power. Note that the great majority of properties had a full 12 months of electricity and gas data in 2012. Slightly fewer properties had a full 12 months of electricity and gas data in 2013. Only about 7 in 10 properties had a full 12 months of data in 2014. CEE does not know the reason for this. It should not have a major impact on measured savings since the cases with less than 12 months of data were reportedly fairly evenly divided between the participant and control groups.

Table 12. Breakdown of Properties by Fuel and Number of Months of Complete Data

Electricity	**year**		
months of data	2012	2013	2014
12	94.8%	92.2%	70.8%
11	1.6%	2.0%	12.1%
10	0.5%	0.7%	3.2%
9	0.5%	0.2%	1.4%
8	0.0%	0.4%	4.6%
7 or fewer	2.5%	4.6%	7.8%
Total (n=561)	100.0%	100.0%	100.0%

Natural Gas	**Year**		
months of data	2012	2013	2014
12	96.6%	92.2%	69.8%
11	1.6%	1.1%	13.4%
10	0.7%	0.9%	4.5%
9	0.0%	0.2%	2.5%
8	0.2%	0.4%	0.5%
7 or fewer	0.9%	5.2%	9.1%
Total (n=553)	100.0%	99.9%	99.9%

Percentages calculated from table of properties or accounts having the specified months of complete data, provided by e-mail by C. Woodson 4/27/2015.

One oddity of the experimental design is that the nominal pre and post years, which are the calendar years 2012 (pre) and 2013 and 2014 (post), do not align with the program start and end dates. According to Bright Power, the pilot project was launched in October 2012 with outreach to the treatment group. The treatment group received access to the EnergyScoreCards tool for two years beginning in the fall of 2012 and ending November 1, 2014. Control group members were contacted during November 2014 to begin access to the service. Thus it would appear that the true program years run from sometime in October 2012 through October 31, 2014, and that the pre-year should have ended at the end of September (or certainly by the end of October) 2012. CEE does not know the reasons for this misalignment.

Bright Power has stated that the activity in October 2012 – January 2013 primarily included notifying portfolios in the sample whether they had been assigned to the treatment group (and would receive access immediately) or to the control group (and would receive access after a two year waiting period) and some early orientation activities for the treatment group. The implication is that they do not think savings are likely to have occurred in the fall of 2012 due to the pilot project and that October-December 2012 can safely be treated as "pre" treatment.

CEE cannot assess what impact this misalignment of program dates and analysis periods (and the corresponding incompleteness of year 2 program data) may have had on the results. If in fact there were savings in the fall of 2012, including that period in the pre year would tend to decrease, rather than increase, measured savings.

5.3. Accounts and Bills Excluded from Analysis

Maximizing the percentage of the original sample retained in analysis is important both to minimize bias and to maximize statistical power. In particular, if properties that opt out of a pilot program are excluded from analysis, the treatment and control groups may no longer be comparable, which may introduce selection bias. Properties that close accounts in a given year (e.g., due to change of ownership) should be dropped from analysis for that year and all subsequent years.

According to Bright Power, "[b]uildings were removed from the final analysis [post year 2 vs. pre year] only if they were deemed invalid, meaning there were less than 235 days of bills for years 2012 or 2014 on any owner-paid energy account." They report that "in some cases data was missing because it had not been received from the utility, wasn't available in the utility's online interface, or because account numbers or login credentials changed and we were unable to obtain updated utility information." Normally account numbers change when an account is closed due to a change of ownership. Changes in login credentials could represent intentional opt-outs, or could have occurred for other reasons, without owner (manager) intent to drop out of the treatment or control group. With the information available it is not possible to quantify the number of accounts excluded due to these various factors.

According to Bright Power, "sixty-seven properties (11.9%) did not have valid data for the baseline and second years in the study, thirty-six in the Control group and thirty-one in the Treatment group." This is a fairly high loss rate. Noting that 235 days of data represents roughly 7.7 months of data (12*235/365) and reviewing Table 12, it appears that most of the drop-out was due to incomplete data for 2014. This is somewhat of a concern since the only savings claimed are for Year 2 of the project, when 10% or more of properties did not have enough data to analyze and only about 70% had a full year of data. The high drop-out rate is unfortunate, but the fact that it is due mostly to incomplete data in the last year of the project suggests, at least, that such a high drop-out rate is not an inevitable aspect of a pilot project in this market, and might be avoided in the future through better alignment of the analysis periods with the period of program operation. A study design that defined the pre year as ending September 30 (or October 31) of 2012 and post year 2 as ending October 31, 2014 might have been able to retain a higher proportion of the initial sample in the analysis, while still ensuring a full year of pre data.

Bright Power reports that, "[i]n order to reduce the impact of incorrect billing data,… each bill is evaluated to see if it strays too far from the fit. If the bill exceeds [a certain (proprietary) number of] standard deviations from the fit, the bill is toggled off, i.e. excluded from the analysis. Once these bills have been excluded, the data is refit, and rechecked for bad bills iteratively. This step is important to reduce the impact of outlier bills produced by utility company estimates or other errors that do not reflect actual changes in consumption. The [multiplier of SD] was chosen through a tuning process on the much larger ESC dataset and was found empirically to exclude a high proportion of bad bills without being overly punitive. Across all properties, this process excluded 1.6% of all bills in the pilot." CEE has not reviewed the bills excluded from analysis to assess whether the stated exclusion criteria are reasonable.

5.4. Data Preparation and Pre-Analysis Equivalency Check

As noted above, maximizing the sample retained for analysis is important both to minimize bias and to maximize statistical power. After reviewing Bright Power's initial round of analysis of the pilot project, CEE offered several suggestions to maximize the number of cases retained in the final analysis:

- Since the primary focus of Bright Power's impact evaluation was the savings from the pre year to post year 2, CEE recommended that only the cases invalid for one of those two years be excluded from analysis, and that cases invalid for 2013 but not 2012 or 2014 be retained in the base year vs. Year 2 analysis. This assumed that cases invalid in 2013 but valid in 2014 had not changed account numbers or opted out in 2013, but merely had missing data due to some data access problem. This change was made in Bright Power's final analyses.

- Twenty-three buildings were missing height category (high-rise/mid-rise/low-rise/garden) in the original Stata data file. Since Bright Power's analysis indicated that the distribution of buildings by height category in the final analysis sample was different for control and treatment buildings, and since they therefore planned to include height category in their regression analysis of savings, CEE recommended that the missing data on height category by filled in by some means to avoid loss of these 23 cases. Bright Power was able to use data on the number of stories in their property data to fill in the missing height category information.
- Building area was missing for three of the buildings, even though the various energy indices were present. Since area is needed to compute aggregate savings from the changes in energy indices, it was recommended that these areas be recovered and included. This was done.
- Year built was missing for 10 properties. Since Bright Power's analyses did not show a significant difference in year built between treatment and control buildings in the final data set, it was recommended that year built be dropped from the regression analysis to avoid exclusion of these 10 cases from analysis. This was done.

CEE identified and pointed out to Bright Power several errors in data preparation (creation of variables within Stata). These were corrected.

Bright Power wisely conducted an equivalency check to determine whether the treatment and control groups remained balanced on various factors that might be related to project outcomes after attrition due to invalid data. They found that, for the most part, the two groups remained well balanced. They appropriately decided to include in the regression analysis those variables that no longer appeared to be well balanced between treatment and control groups.

Bright Power computed correlation coefficients for the building characteristics and other data available for potential use in regression analysis, and appropriately decided to include only one of any pair of highly correlated variables in the regression, in order to reduce potential problems with multi-collinearity.

5.5. Specification of Analysis Models and Reporting of Analysis Results

5.5.1 Specification of Analysis Models

The SEE Action Network[10] report on recommended evaluation methods for behavioral energy efficiency programs delivered to single family homes made several recommendations for optimal analysis of such programs. These programs have small average savings, which can be difficult to measure without both large samples and analysis models that estimate savings as precisely as possible.[11] Since the EnergyScoreCards service also seeks to influence behavior -- through the similar mechanism of providing feedback that allows the party responsible for energy costs to compare their energy use to normative energy use -- savings from the EnergyScoreCards service were also expected to be fairly small in

[10] Op. cit.

[11] That is, with the smallest possible standard errors of the mean savings estimates.

percentage terms. Therefore, the SEE Action Network recommendations regarding models that can increase precision in randomized controlled trials (RCTs) are relevant to this pilot project.

To increase the precision of savings estimates in RCTs, SEE Action Network recommended that these evaluations use panel data models, so that the analysis is run directly on monthly data, rather than aggregated models in which the analysis is run on annual totals. They further recommended that the analysis models compare the energy saved by the treatment and control groups, rather than the energy use of the two groups, and that a few logically chosen control variables – variables that may help to explain patterns of energy use not related to the program --be included in the analysis.

A further very important consideration in the specification of analysis models for this project (not discussed by the SEE Action Network because it was not applicable to the program designs they were covering) is that this sample is a stratified, cluster-randomized sample, rather than a simple random sample. If clustered data are analyzed as if they were from a simple random sample, the precision of the savings estimates and their statistical significance will be substantially overestimated. Therefore it is important to use an analysis model that takes into account the clustered nature of the data.

In its first year evaluation report, Bright Power appropriately compared the change in energy use of the participant and control groups rather than energy use itself, but did not take into account the stratified, cluster-randomized sample design. Nor did they use a panel data model or include control variables in the analysis.

CEE discussed these issues extensively with Bright Power staff prior to their analysis of year 2 data. CEE recommended that Bright Power use a multilevel (aka hierarchical) model to conduct the final analysis. This would take into account the clustered structure of the data and enable the direct use of monthly data in a panel data framework, as well as the inclusion of control variables at both the property and portfolio levels.

Bright Power chose to address the issue of clustering in the data by conducting regression analysis with cluster-robust standard errors, rather than by using a multilevel model. This approach should produce results with correct levels of statistical significance, if it fully accounts for the structure of the data and if no cluster-level variables (portfolio characteristics, as opposed to building characteristics) are included in the model. Both Bright Power and CEE forgot about the super-clusters and stratification included in the sample design until well into the analysis work. CEE conducted some supplementary analysis to assess whether including those factors in the analysis model would materially change the findings, and concluded it did not. An updated version of this analysis, using the last version of the dataset used by Bright Power in analysis, is presented below.

For the entire pilot project sample, using cluster-robust standard errors with superportfolio3 as the cluster variable does not change the estimate, but, increases the p value slightly, as shown in the first regression below. Adding stratum3 as a control variable, as shown in the second regression below, changes the savings estimate slightly from 1.326 to 1.337 kBtu/sf-yr, and also decreases the p value slightly, even though it causes a loss of some degrees of freedom, because stratum3 has a small beneficial effect as a control variable.

```
.   regress   energyChange   i.treated   c.eui2012   ib2.height_category   c.squareFeet   i.subsidy,
vce(cluster superportfolio3)

Linear regression                                      Number of obs =     492
                                                       F(  6,    76) =    4.61
                                                       Prob > F      =  0.0005
                                                       R-squared     =  0.0943
                                                       Root MSE      =  6.2457

                          (Std. Err. adjusted for 77 clusters in superportfolio3)
---------------------------------------------------------------------------------
                |               Robust
   energyChange |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
        treated |
   Participant  |  -1.325679   .8699344    -1.52   0.132    -3.058304    .4069453
        eui2012 |  -.0517434   .0207153    -2.50   0.015    -.0930016   -.0104852
                |
height_category |
           Hig  |  -5.363625   1.655685    -3.24   0.002    -8.661207   -2.066043
           Mid  |   1.426426   1.599113     0.89   0.375    -1.758485    4.611336
                |
     squareFeet |  -1.93e-07   3.34e-06    -0.06   0.954    -6.85e-06    6.46e-06
      1.subsidy |  -.3550443   .8429941    -0.42   0.675    -2.034013    1.323924
          _cons |   3.759183   1.270575     2.96   0.004     1.228614    6.289753
---------------------------------------------------------------------------------
```

```
. regress energyChange i.treated c.eui2012 ib2.height_category c.squareFeet i.subsidy i.stratum3,
vce(cluster superportfolio3)

Linear regression                                      Number of obs =     492
                                                       F(  8,    76) =    3.69
                                                       Prob > F      =  0.0011
                                                       R-squared     =  0.1015
                                                       Root MSE      =  6.2334

                          (Std. Err. adjusted for 77 clusters in superportfolio3)
---------------------------------------------------------------------------------
                |               Robust
   energyChange |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+----------------------------------------------------------------
        treated |
   Participant  |   -1.33733     .87201    -1.53   0.129    -3.074089     .399428
        eui2012 |  -.0468997   .0211748    -2.21   0.030    -.0890729   -.0047265
                |
height_category |
           Hig  |  -5.459012   1.627138    -3.35   0.001    -8.699738   -2.218286
           Mid  |   1.265913    1.55324     0.82   0.418    -1.827632    4.359459
                |
     squareFeet |   6.09e-07   3.51e-06     0.17   0.863    -6.39e-06    7.61e-06
      1.subsidy |  -.1393092   .8766842    -0.16   0.874    -1.885377    1.606759
                |
       stratum3 |
             2  |  -1.583131   1.611042    -0.98   0.329    -4.791799    1.625537
             3  |    .055653   .9105275     0.06   0.951     -1.75782    1.869126
                |
          _cons |   3.595433   1.356672     2.65   0.010     .8933859    6.297479
---------------------------------------------------------------------------------
```

For the group with payment code = (O)OOO, it is also the case that taking account of the superclusters does not change the estimate of savings, as shown in the first regression below. The p value is also virtually unchanged. Taking stratum into account as well, as shown in the second regression, increases the savings estimate moderately from 4.188 to 5.076 kBtu/sf-yr and improves (decreases) the p value, because in this subset stratum is more effective as a control variable. Given the small number of properties in the (O)OOO group (84), this analysis would have to be reviewed more closely, but for the present purposes the analysis does show that, had Bright Power taken supercluster and stratum into account in the analysis, it would not have changed the conclusions of the impact evaluation to a great extent. The entire sample is still shown not to have realized significant savings, and the (O)OOO sample is still shown to have done so.

```
. regress energyChange i.treated c.eui2012 ib2.height_category c.squareFeet i.subsidy if payOOOO,
vce(cluster superportfolio3)

Linear regression                                      Number of obs =      84
                                                       F(  6,    33) =    1.80
                                                       Prob > F      =  0.1297
                                                       R-squared     =  0.1737
                                                       Root MSE      =  7.8289

                          (Std. Err. adjusted for 34 clusters in superportfolio3)
--------------------------------------------------------------------------------
               |               Robust
  energyChange |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
       treated |
  Participant  |  -4.187751   1.545764    -2.71   0.011    -7.332633    -1.04287
       eui2012 |  -.0394077   .0516841    -0.76   0.451    -.1445597    .0657443
               |
height_category |
          Hig  |  -7.199416   4.150689    -1.73   0.092    -15.64406    1.245225
          Mid  |   .9008279   3.675845     0.25   0.808    -6.577734     8.37939
               |
    squareFeet |   9.64e-06   .0000198     0.49   0.630    -.0000307      .00005
     1.subsidy |   1.687056   3.465707     0.49   0.630    -5.363979    8.738091
         _cons |   1.389786   5.653601     0.25   0.807    -10.11255    12.89213
--------------------------------------------------------------------------------


. regress energyChange i.treated c.eui2012 ib2.height_category c.squareFeet i.subsidy i.stratum3
if payOOOO, vce(cluster superportfolio3)

Linear regression                                      Number of obs =      84
                                                       F(  7,    33) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.2159
                                                       Root MSE      =  7.7273

                          (Std. Err. adjusted for 34 clusters in superportfolio3)
--------------------------------------------------------------------------------
               |               Robust
  energyChange |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
       treated |
  Participant  |  -5.076071   1.751471    -2.90   0.007    -8.639466   -1.512676
       eui2012 |  -.0421575   .0517089    -0.82   0.421    -.1473601     .063045
               |
height_category |
          Hig  |  -8.639242   4.229147    -2.04   0.049    -17.24351   -.0349779
          Mid  |   1.707868    4.51662     0.38   0.708    -7.481265      10.897
               |
    squareFeet |   .0000109   .0000179     0.61   0.549    -.0000256    .0000473
     1.subsidy |   1.407585   3.010426     0.47   0.643    -4.717172    7.532342
               |
      stratum3 |
            2  |  -2.789899    2.84424    -0.98   0.334    -8.576549    2.996751
            3  |  -14.23329    4.14156    -3.44   0.002    -22.65936    -5.80722
               |
         _cons |   4.492048   6.471077     0.69   0.492    -8.673457    17.65755
--------------------------------------------------------------------------------
```

Bright Power did not consider it feasible to use monthly data directly in a panel data analysis, because there were often multiple energy accounts at each property and these typically were not on the same billing cycle, so the meters were not read on the same day. Rather, they preferred to normalize the data for weather and produce annual, weather-normalized consumption values prior to computing various indices and year to year changes in indices, with the latter forming the dependent variables in their analysis models. This choice may well be necessitated by the data (which CEE has not reviewed) and only affects the precision of savings estimates (i.e., it does not produce biased estimates), so it is acceptable.

Weather-normalized monthly values could not be used in a panel data model in any case because, after normalization, they are no longer independent monthly readings but rather monthly values generated by a normalized energy use model. However, it is possible to use a panel data model for the annual totals, since they are weather-normalized separately. In such a model, individual years rather than months would be used as the time steps at which data are available. An annual panel data model obviously has far fewer observations than a monthly panel data model, and much less potential for gains in precision, but it is worth exploring because it allows the use of fixed effects regression. This is a technique that allows the analysis to focus on the changes that occur over time, rather than the "fixed" differences that exist between properties. Indeed, any variable that does not change from the pre to the post period (such as height category or building area) drops out of the regression. This approach reduces the risk of confounding building to building differences with pilot project impacts. A fixed effect model regresses the difference of the dependent variable from its building-level mean at each point in time on the difference of the independent variables from their building-level means.[12] In so doing it effectively controls for building-to-building variation in mean energy consumption (or EUI, in this case), regardless of the factors that may account for that variation. Each building's mean consumption (or mean EUI) over the three year period is simply considered a "fixed effect" for that building. CEE conducted a simple annual fixed-effects analysis for owner EUI to see whether it suggested results substantively different from those obtained using Bright Power's analysis model, and concluded that it did not. An updated version of this analysis, using the last version of the dataset used by Bright Power in analysis, is presented below.

For the entire group, the fixed effects regression (xtreg, fe) using superportfolio3 as the cluster variable produces an estimate of savings in 2014 (Participant#2014) of 1.14 kBtu/sf-yr, as shown in the first regression below, which is not significant (p=0.198). This is very similar to the result from Bright Power's regression analysis. For the group with payment code = (O)OOO (second regression below), the fixed effects regression shows savings in 2014 of 4.73 kBtu/sf-yr, with p = 0.03. This again is very similar to the result from Bright Power's regression analysis.

```
. xtreg eui i.treated##i.year, fe vce(cluster superportfolio3) allbaselevels

note: 2.treated omitted because of collinearity
```

[12] Stata adds the grand means across all households back into the regression, which affects the constant in the regression but not the coefficients (Gould, W. 1997, updated 2013. "How can there be an intercept in the fixed-effects model estimated by xtreg, fe?" Stata FAQ. Available at http://www.stata.com/support/faqs/statistics/intercept-in-fixed-effects-model/).

```
Fixed-effects (within) regression               Number of obs      =      1479
Group variable: origID                          Number of groups   =       493

R-sq:  within  = 0.0129                         Obs per group: min =         3
       between = 0.0065                                        avg =       3.0
       overall = 0.0013                                        max =         3

                                                F(4,76)            =      1.16
corr(u_i, Xb)  = 0.0175                         Prob > F           =    0.3367

                            (Std. Err. adjusted for 77 clusters in superportfolio3)
----------------------------------------------------------------------------------
                  |               Robust
              eui |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+---------------------------------------------------------------
          treated |
         Control  |          0  (base)
     Participant  |          0  (omitted)
                  |
             year |
            2012  |          0  (base)
            2013  |  -1.110598   .7837925    -1.42   0.161    -2.671656      .45046
            2014  |    .432691   .3997079     1.08   0.282    -.3633964    1.228778
                  |
     treated#year |
    Control#2012  |          0  (base)
    Control#2013  |          0  (base)
    Control#2014  |          0  (base)
Participant#2012  |          0  (base)
Participant#2013  |   .4228016   1.014108     0.42   0.678    -1.596969    2.442572
Participant#2014  |  -1.143161   .8793945    -1.30   0.198    -2.894627    .6083044
                  |
            _cons |   57.15739   .2642635   216.29   0.000     56.63107    57.68372
------------------+---------------------------------------------------------------
          sigma_u |  28.652563
          sigma_e |  5.4805369
              rho |  .96470495   (fraction of variance due to u_i)
----------------------------------------------------------------------------------

. xtreg eui i.treated##i.year if pay0000, fe vce(cluster superportfolio3) allbaselevels
note: 2.treated omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =       252
Group variable: origID                          Number of groups   =        84

R-sq:  within  = 0.1052                         Obs per group: min =         3
       between = 0.0025                                        avg =       3.0
       overall = 0.0000                                        max =         3

                                                F(4,33)            =      2.24
corr(u_i, Xb)  = -0.0569                        Prob > F           =    0.0855

                            (Std. Err. adjusted for 34 clusters in superportfolio3)
----------------------------------------------------------------------------------
                  |               Robust
              eui |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+---------------------------------------------------------------
          treated |
         Control  |          0  (base)
     Participant  |          0  (omitted)
                  |
             year |
            2012  |          0  (base)
            2013  |    .120079   .5528839     0.22   0.829    -1.004772     1.24493
            2014  |  -.5009076   .8366427    -0.60   0.553     -2.20307    1.201255
                  |
     treated#year |
    Control#2012  |          0  (base)
    Control#2013  |          0  (base)
    Control#2014  |          0  (base)
```

```
Participant#2012  |          0  (base)
Participant#2013  |  -3.437891   1.762608    -1.95   0.060    -7.023943    .1481614
Participant#2014  |  -4.732054   2.084938    -2.27   0.030    -8.973893   -.4902149
                  |
            _cons |   82.49142   .5825813   141.60   0.000     81.30615    83.67669
------------------+----------------------------------------------------------------
          sigma_u |  36.103427
          sigma_e |  5.7576951
              rho |  .97519763   (fraction of variance due to u_i)
```

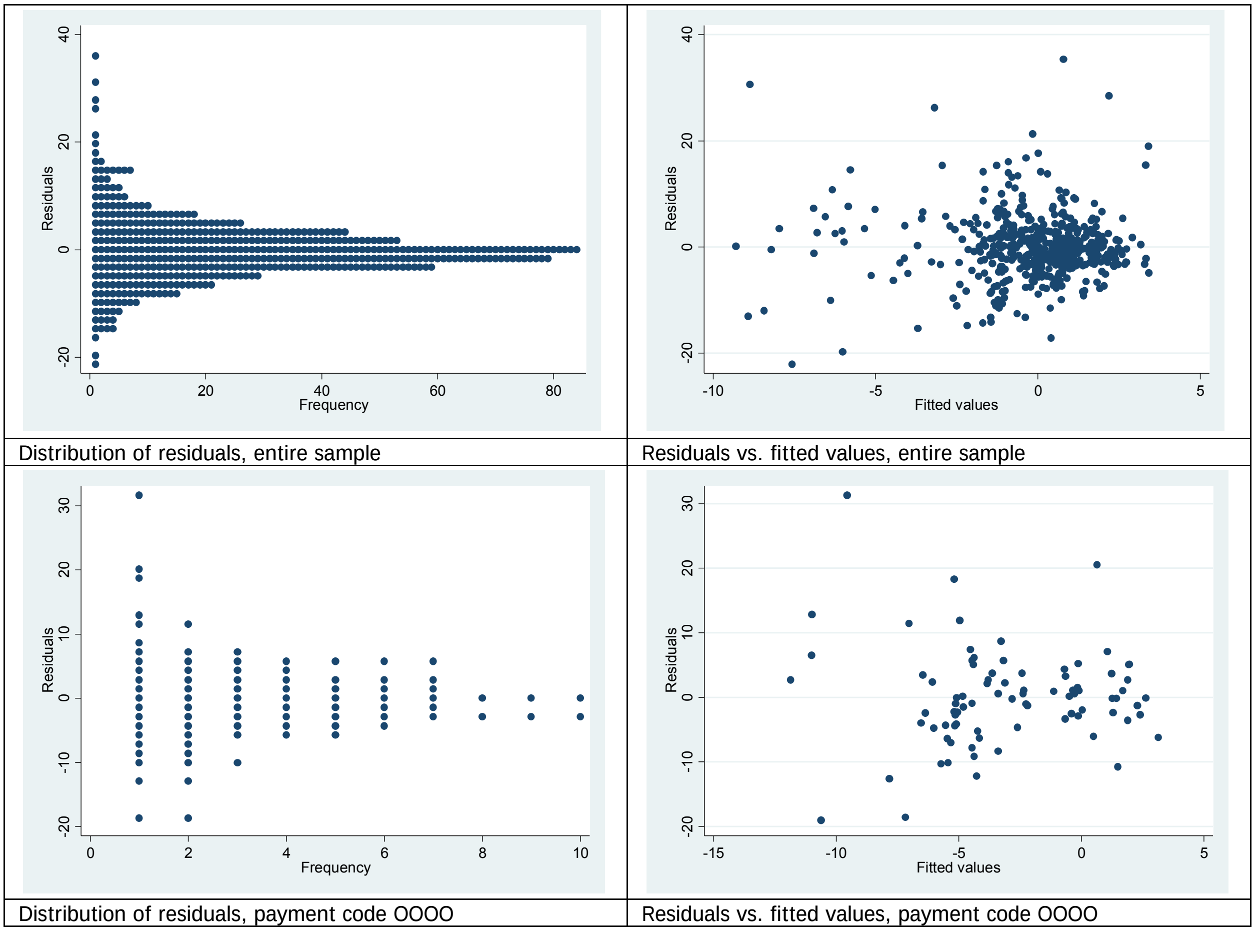
Bright Power did include control variables in its regression models. CEE provided input on this, recommending that variables that did not differ between the participant and control groups be omitted unless there was a specific reason for including them. If such variables have missing values, this decreases the sample size, and may bias the outcome if the cases with missing values differ from those without. In addition, including these variables uses up degrees of freedom, and with only about 88 clusters in the final data set, Bright Power needed all the degrees of freedom it could reasonably preserve. Finally, including such variables was unlikely to substantially increase the R^2. Bright Power followed this recommendation.

CEE also provided input to on several other aspects of Bright Power's analysis models. In their first round of analysis for the final impact evaluation, Bright Power included analysis of the impact of the number of logins on savings. The thought was that owners (managers) who logged in more times might save more. The analyses as structured had the effect of lumping participants who did not log in at all in with the control group, so that it confounded logins with participation. In addition, some of the analyses had interaction terms in the regression (e.g., logins interacted with payment codes) without including the main terms. This formulation will not produce correct results. CEE explained and demonstrated the problems with these analyses to Bright Power staff. Corrected analyses of the impact of logins were not included in the final report, perhaps for lack of time to follow up in revising these analyses.

CEE suggested that, if savings were not significant for the entire pilot group, Bright Power might want to examine subsets of the data that could be expected to have less variability in savings, making savings more discernible. For example, analyzing changes in heating and hot water energy use for all accounts where owners pay for this use might be expected to reduce year to year variability due to things like electric base load use and provide a large subset with consistent owner motivation, and therefore produce more consistent savings that could be more discernible with the available sample size. CEE also warned Bright Power to be careful about drawing conclusions from patterns that were not statistically significant, noting that the first year report appeared to have some discussion of patterns of savings across building "grades" (Bright Power's A-D ratings based on relative energy use) that were not supported by statistical tests. In the final report, Bright Power did include results of appropriate statistical analyses of savings within relevant subsets of participants.

A final aspect of the analysis models that Bright Power did not address is the assumption inherent in regression that the residuals are normally distributed and have constant variance. Given the wide range of building sizes and types and building energy use, CEE felt that it would be wise to verify that these assumptions were reasonable, and therefore conducted some quick checks. An updated version of these checks, using the last version of the dataset and regression models used by Bright Power in analysis, is presented in the graphs below. The dot plots on the left show that the residuals for the regressions with energyChange (eui2014-eui2012) as the dependent variable are not skewed and are at least grossly normal. The scatter plots on the right show that these same residuals are distributed roughly consistently as a function of the fitted values; that is, they appear to have roughly uniform variance.

Figure 2. Plots to check approximate normality (left) and homoscedasticity (right) of regression residuals



Distribution of residuals, entire sample

Residuals vs. fitted values, entire sample

Distribution of residuals, payment code OOOO

Residuals vs. fitted values, payment code OOOO

5.5.2. Reporting of Analysis Results

Bright Power's first year evaluation report provided very few details on the statistical analysis conducted. CEE therefore provided considerable guidance to Bright Power on the reporting of results, both in checklists provided before work began and through review of two drafts of the report. These recommendations included the following items:

- Show the specific statistical methods used to test for significant program savings and the specification and output of those tests in an appendix, and describe them thoroughly in the text.
- Do not present t tests or regressions that do not use cluster-robust standard errors in the report.
- Do not present results with p values > 0.05 as statistically significant. Results with p values between 0.05 and 0.1 can at most be described as marginally significant.
- Include both year 1 and year 2 results in the report, and be clear throughout that the savings reported are only for the second year.

- Report all savings as a percentage of pre-program energy use and as aggregate pilot project savings (total kBtu/year, total gallons/year, etc.,) as well as in the form of average changes in EUI and other indices that come directly from the regression results. This will give the reader an understanding of the savings relative to total building energy use, and of the total pilot project savings.
- When reporting that heating savings are a majority of total savings, calculate and present the actual percentage, since the indices are in different units and cannot be directly compared by the reader.
- Estimate savings for electricity and gas separately, as well as total energy savings. Savings by fuel type are of interest to the utilities and are required to compute dollar savings.

5.6. Methods Used to Address Double-Counting of Savings

Programs that are designed to induce behavioral change by providing customers with normative information about their energy use may have the effect of causing the participant group to use utility rebate programs more than the control group. This increased uptake of rebates is a clear benefit of the feedback (benchmarking) program. However, when such a program is included in utility energy efficiency portfolios, it is necessary to avoid double-counting the savings that are jointly due to the feedback program and the rebate program. In Minnesota, the practice thus far has been to assign the savings to the rebate program. In CEE's opinion, this should be accompanied by assigning some of the costs of the feedback program to the rebate program, since the feedback program is operating as a de facto marketing mechanism for that program. Assigning all of the joint savings but none of the feedback program costs to the rebate program unfairly lowers the apparent feedback program cost-effectiveness. Nevertheless, the practice thus far in Minnesota has not assigned any of the costs associated with achieving the increased uptake of rebates to the rebate program. The cost-effectiveness of the feedback or benchmarking program is thus taken to be (total minus joint savings)/total costs.

Prior to the start of evaluation work, CEE acquainted Bright Power with these practices and advised them to begin early in requesting rebate program data from the utilities. CEE also recommended that these joint savings be determined using a difference-in-differences calculation, i.e., taking the difference between how much the use of rebates increased among participants from the pre year to the post years and how much the use of rebates increased among controls from the pre year to the post years. CEE also pointed out that rebates given in the first year for measures with a longer life are also responsible for joint savings in the second year, and these, too, must be deducted. CEE also advised Bright Power, based on CEE's past experience, that it would be important to review the deemed savings data provided by the utilities to make sure it appeared reasonable before using it, since sometimes utilities' rebate databases can get corrupted.

Bright Power did include analysis of joint savings in their final evaluation. However, the analysis did not follow CEE's recommendations in several respects. First, Bright Power did not obtain information from the utilities on rebate use in the baseline year, so could not conduct a difference-in-differences analysis. Rather, they simply took the difference in use of rebates between participants and controls in the post years and used it as the incremental impact. This is not as accurate, but since the incremental rebates accounted for only 16% of claimed total ESC project electric savings and 0% of gas savings, the inaccuracy is probably not large. Bright Power also did not obtain and review building-level data, as CEE expected, but rather total pilot project numbers complied by the utilities. This created an unanticipated problem in that ESC project savings were only claimed for the (O)OOO group, but incremental use of utility rebates specific to participants in the (O)OOO group could not be ascertained because the rebate data provided by the utilities were aggregated totals. There is no way to know whether incremental use of rebates in the (O)OOO group was higher or lower than the average across all payment codes in the pilot. Finally, we do not know whether or not Bright Power determined the incremental difference in rebate use separately for years 1 and 2, and discounted the joint savings for long-lived measures implemented in year 1 in both years 1 and 2. Since the joint savings appear to be a relatively small part of total pilot project savings, the error due to these issues is probably not terribly large.

5.7. Cost-Effectiveness

Bright Power did not conduct the types of cost-effectiveness analysis typical for utility programs, but rather a simplified analysis of dollars saved by the owners (at retail rates) per dollar spent to deliver the program.

CEE provided multiple rounds of review and comment on this analysis. Among the recommendations were:

- Recognize in analysis of overall project cost-effectiveness that there were no statistically significant savings in the first year.
- Determine the fraction of savings jointly attributable to the ESC service and utility rebate programs. Following current practice in Minnesota, exclude these joint savings from calculation of cost-effectiveness.
- Conduct and include in the report separate analyses of electricity and gas savings to allow savings in dollars to be calculated.
- Recognize that both treatment and control buildings incur costs but only treatment buildings realize savings.
- Include all costs likely to be incurred in a full scale program. These include not just the deployment phase of the project but also the management, program design, training, marketing, data collection, and evaluation.

The first draft of the report reviewed by CEE reported savings of $31/year for every $1/year spent. The last draft CEE saw before preparation of this audit report reported savings of $2.15/year for every $1/year spent.

Bright Power's extrapolation of the pilot project results for master-metered buildings (payment code (O)OOO) to estimate cost-effectiveness for a 10 year program is clearly hypothetical, since the two year pilot project was not long enough to provide a basis for estimating the persistence of savings from the ESC service. However, the assumptions are clearly stated and the results are not claimed to represent actual pilot project cost-effectiveness.

5.8. Post-Hoc Statistical Power Analysis

As noted in Section 2, it is useful as part of project planning to conduct a statistical power analysis. This provides an estimate of the sample size required to detect savings of the anticipated magnitude. In the case of the current project, the sample size was driven by budget constraints rather than strict statistical power considerations. However, it is still useful to conduct a power analysis at the conclusion of the project, to provide some insight into possible reasons why significant savings were not observed, and to provide some guidance for future projects.

CEE recommended at the start of the evaluation phase that Bright Power conduct a power analysis. We subsequently pointed out that their initial power analysis did not take the clustered sample design into account, identified a Stata add-in that could be used for cluster sample power analysis, provided relevant reference articles and conducted an initial analysis to illustrate methods to arrive at the parameter values required as input for the analysis. Bright Power subsequently researched this issue further and conducted a revised analysis.

As shown in Bright Power’s cluster sample size analysis command in Stata,

```
clustersampsi, samplesize mu1(0) mu2(-1.3) sd1(6.1) sd2(6.9) m(5) rho(0.14726)
size_cv(0.9) base_correl(0.3)
```

the required sample size will depend on the mean change for the control group and participant group (mu1 and mu2), the standard deviation of the change for those two groups (sd1 and sd2), the average cluster size (m), the coefficient of variation (standard deviation/mean) of the cluster size (size_cv), and the intraclass correlation coefficient (rho) within the clusters (portfolios). All of the parameter values in the above command are the actual observed values from analysis dataset for the pilot project.

With this analysis Bright Power determined that a sample of 790 buildings per arm (participant and control arms) or 1580 total, in 316 clusters (portfolios) would be required to detect as statistically significant a change of -1.3 kBtu/sf-yr.[13] This is a large number of buildings to recruit and provide benchmarking for, and raises the question of how a smaller sample could be used while still obtaining a statistically significant result. One option is to use unclustered data, i.e., to recruit only one building from each management company. The Stata output shows that this would require 359 buildings per arm or 718 buildings total, larger than the sample size in the current study but much smaller than a 1580 building sample. However, the current pilot project recruited only 93 owners/management companies. Recruiting 718 companies could be much more difficult. The only parameters that the experimenter can control in the calculation of cluster sample size are the average number of buildings per cluster, m, and the variation in cluster size (size_cv). In the current project, if the average final cluster size had been kept the same as it was (5) but the maximum number of buildings per cluster had been reduced from 24 to 12, the size_cv would have dropped to 0.80 and the required sample size would have dropped from 1580 buildings in 316 portfolios to 1490 buildings in 298 portfolios, all other things being equal. If the maximum cluster size had been further reduced to 8, while still keeping the average size at 5, the size_cv would have dropped to 0.74 and the required sample size to 1440 buildings in 288 portfolios. If the

[13] The clustersampsi analysis used the typical default power criterion of 80% and significance level of 5%.

average cluster size (m) had been reduced from 5 to 4 with a maximum size of 8, the size_cv would have dropped to about 0.65 and the number of buildings required would have been reduced to 1224 buildings in 306 portfolios. These variations are worth investigating for future planning, to assess the most cost-effective sample size and recruitment strategy in a future pilot or full scale program.

It should be noted that the `clustersampsi` calculations shown above do not take into account any loss of cases due to data problems. Therefore in a future project the required sample sizes determined from such calculations should be inflated by a reasonable percentage to account for the anticipated number of cases in the initial sample that will not be usable due in the final analysis due to such data problems.

In addition to estimating the sample size required to obtain a statistically significant result given the observed (non-significant) savings, it is also possible in retrospect to determine the smallest savings that would have been detectable (significant) with the actual sample size and sample design. This can be done in Stata using the "detectabledifference" option with the `clustersampsi` command, as shown below:[14]

```
. clustersampsi, detectabledifference mu1(0) sd1(6.5) m(5) k(45) rho(0.14726) size_cv(0.9)
base_correl(0.3)

Detectable difference calculation for two sample comparison of means (using normal
approximations)and assuming equal standard deviations.

For the user specified parameters:
mean 1:                                                         0.00
standard deviation 1:                                           6.50
significance level:                                             0.05
power:                                                          0.80
baseline measures adjustment (correlation):                     0.30
average cluster size:                                           5
number of clusters per arm:                                     45
coefficient of variation (of cluster sizes):                    0.90
intra cluster correlation (ICC):                                0.15
clustersampsi estimated parameters:
Firstly, under individual randomisation:
detectable difference:                                          1.64
If, trying to detect an increasing outcome then:
corresponding mean 2:                                           1.64
If, trying to detect a decreasing outcome then:
corresponding mean 2:                                           -1.64
Then, allowing for cluster randomisation:
design effect:                                                  2.19
detectable difference:                                          2.45
If, trying to detect an increasing outcome then:
```

[14] Since only one standard deviation is entered as an input parameter, a value midway between the standard deviations of the sample and control groups was used.

```
corresponding mean 2:                                           2.45
If, trying to detect a decreasing outcome then:
corresponding mean 2:                                           -2.45
Note: standard deviations assumed equivalent in both arms.
```

This analysis shows that the smallest difference that could have been detected with the sample size and sample design used in the pilot project was 2.45 kBtu/sf-yr. Given that the average 2012 owner EUI was 56.98 kBtu/sf-yr, this means that an average savings of 4.3% would have been required for it to be detectable with the same size and sample design used. This is within the range of savings (3 to 5%) that Bright Power thought they might observe during project planning in fall 2011, but obviously is higher than was actually obtained in this particular pilot.