The Devil's in the Details:
What is the Effect of Methodological Decisions on Estimated Savings?

Pace Goodman, Navigant, Moscow, ID
Jes Rivas, Navigant, Vancouver, WA
Marilla Yaggie, Navigant, Boulder, CO
Lauren Gage, Apex Analytics, Coeur d'Alene, ID

ABSTRACT

How often have you seen estimates of savings achieved through an energy efficiency program vary from one residential consumption analysis study to the next, while the evaluation method appears consistent? How do you decide which results are most accurate? And how can you compare results across these studies?

Energy efficiency programs use rebates and other mechanisms to encourage cost-effective energy efficiency interventions. These programs are evaluated to ensure they are cost effective and achieving the expected impact on energy consumption, which requires consistent, reliable and accurate results. However, the methods used for evaluating programs using consumption data vary. Although efficiency evaluation has decades of industry experience including several guidelines for consumption analysis, such as the Department of Energy's Uniform Methods Project (UMP), there is still substantive variation in consumption analysis methods and results. This variation is partially driven by academia, where active research is rapidly evolving methods, and partially by analyst subjectivity, where evaluators make project-specific judgements on methods.

This paper presents an overview of some of the variation and vagueness in industry documents and provides an indication of the potential impact on results from certain analysis decisions. It investigates the level of variation in savings estimates when savings from a single data set are calculated using different model types, comparison groups and data filters. This work informs evaluators, researchers, policy makers and program administrators as to the relative importance of certain analysis decisions, such as model selection, comparison group development and data filtering.

Introduction

This paper focuses on residential consumption analysis (also referred to as billing analysis) for estimating energy savings achieved through the implementation of an energy efficient intervention, such as installing a more efficient device. The goals of this paper are to a) identify common differences in evaluation practices, guidance documents, and academic literature and b) provide a spectrum of results that help to indicate the impact and variability caused by each inconsistency. Before discussing findings and methods, the authors provide a brief background on the following topics:

- Consumption analysis
- Status of consumption analysis standardization

Consumption Analysis Overview

Consumption analysis is used to estimate the overall energy impact of the measures delivered by a program, and it is often used to determine the portion of that impact that should be attributed to the program. One or more regression models are used to analyze utility billing data. The models themselves vary depending on research needs and available data, but the basic objective is always the same:

empirically estimate changes in energy consumption that are associated with program interventions (Gage 2015).

In the residential sector, consumption data analysis is typically used to evaluate the following energy efficiency measures:

- Whole-building retrofits with multiple measures, such as combinations of insulation, air-sealing and duct sealing
- Heating and cooling measures, including HVAC tune-ups, duct sealing, heat pump conversions, ductless heat pumps and thermostat replacements
- Behavioral measures, like home energy reports and certain advanced thermostat features

This type of analysis is only applicable to measures with pre-existing baselines (i.e., as-is conditions, as opposed to estimating savings against a federal standard efficiency or a code baseline, for example), and when the magnitude of savings and the sample size are large enough to yield statistically significant results.

Consumption data analysis can be broken into three main steps: data processing, control or comparison group development, and modeling (typically in the form of multivariate regression models).

Status of Consumption Analysis Standardization

The evaluation industry has substantial experience in using consumption analysis to estimate energy savings for specific programs (Agnew 2013; Apex 2016; Cadmus 2012; Cadmus 2015; Greco 2015; Efficiency Valuation Organization 2012; Kelsven 2016; Olig 2016; Patterson 2016; Schiller 2012; Stewart 2015; Violette 2014). Additionally, academic studies and industry studies continue to publish informative research that could, and perhaps should, influence evaluation methods (Alcott 2012; Ho 2007).

In recognition of the potential for variation in methods as the industry learns and grows, industry guidance documents provide recommendations for methods and data management. In 1997, the first edition of the International Performance Measurement and Verification Protocol (IPMVP) provided recommended methods to quantify savings in an effort to encourage and justify investment in energy conservation. In 2013, the Department of Energy started releasing more specific recommended analysis guidelines for evaluating demand side management (DSM) programs through the Uniform Methods Project (UMP). The UMP now provides over 20 guidelines, which they refer to as protocols. Chapters 8, 17 and 23 all discuss consumption analysis.

In reviewing these documents, it is common to find variation in model selection, comparison group development and filtering approaches to specify the inclusion/exclusion of sites from the analysis data set. For example, the three UMP chapters that discuss consumption analysis cover at least four different model types, from which there can be additional variation based on the inclusion or exclusion of independent variables. These chapters also discuss developing a comparison group using future participants' pre-upgrade usage data, past participants' post-upgrade usage data and non-participant usage data. Filtering is only discussed in detail in UMP Chapter 8, but in the industry, filtering is implemented in a variety of ways.

Fundamentally, the goal of evaluation is to estimate savings accurately. The analysis decisions that drive the greatest variation in savings can be interpreted as the most important inconsistencies for the industry to confront in order to achieve more reliable and more accurate estimates of savings.

Available Data

In this section the authors discuss the available data for this work. This data includes billing data, weather data and a reporting database for a set of efficiency measures.

Efficiency measures: The focus of this effort includes the following measures:

- Residential duct sealing
- Residential ductless heat pumps (DHP) replacing electric forced air furnaces (eFAF)
- Residential windows
- Residential insulation

Billing data: The authors collected billing data from 22 utilities in the Northwest, including Washington, Oregon, Idaho, and Montana. For these utilities, the authors requested a census of participant billing data (from two years prior and everything following the measure installation) for all evaluation measure groups.

Weather data: The authors gathered two sets of weather data (actual and typical year) that were used to control for varying weather conditions in the analysis and to calculate weather-normalized savings (i.e., savings during a typical meteorological year (TMY)). The authors used participants' zip codes to match their locations to weather stations in the National Oceanic and Atmospheric Administration's National Climatic Data Center database of historical weather data. The authors ensured that each weather station had adequate weather data and available TMY data from the National Renewable Energy Lab's National Solar Radiation Database.

Reporting database: Navigant also analyzed program data stored in a database that includes all the utilities incentivized and rebated measures. This database was queried to identify all efficiency measures installed at study participant sites during the evaluation period and when those measures were installed. The authors used this data to ensure their billing analysis did not double count savings and to ensure that their analysis accounted for real world measure interaction. Double counting in this context refers to attributing all energy savings at a site, which often come from multiple measures, to a single measure. Knowing all the measures installed at each study participant site enables the authors to attribute savings to single measures more accurately.

Methods & Results

The following sections describe the methods and results when using different model types, comparison groups and data filters. Navigant worked with regional stakeholders through the Regional Technical Forum's Statistical Sub-committee[1] to identify the most important models to be run and filters to investigate. This investigation is not exhaustive, but is intended to provide an indication of the importance of these topics.

Furthermore, these results are draft and this work is ongoing. The results[2] presented here are aggregated and anonymized to avoid sharing any confidential results or information.

Model Types

Model types are used to derive savings estimates from consumption data. Almost all consumption data models involve multivariate linear regressions. However, within this framework there can still be variation. Models can use different specification (i.e., the independent variables used to model consumption), staging (i.e., whether models are applied first at the site level and then to the population), as well as many other areas of variation that can lead to inconsistent results. Although models can vary with the inclusion or exclusion of certain independent variables or covariates, the authors expect to see

[1] The RTF is a technical advisory committee to the Northwest Power and Conservation Council established in 1999 to develop standards to verify and evaluate energy efficiency savings - https://rtf.nwcouncil.org/

[2] Savings are presented with a fixed y-axis range.

larger variation between model types (i.e., the framework of the model). For this effort, the authors focused on three modeling types (below):

- Lagged Dependent Variable or Post-Only, see Alcott (2012) or for more details
- Two-stage or variable base degree day (VBDD) approach, see Agnew (2013) for more details
- Difference-in-difference, see Violette (2014) or Stewart (2015) for more details

Figure 1 shows the range of savings for different model types across four measures. These results show that model selection can have substantive impacts on results. The results are sometimes statistically different and there is not a consistent trend, where one model provides consistently lower/higher savings than others.

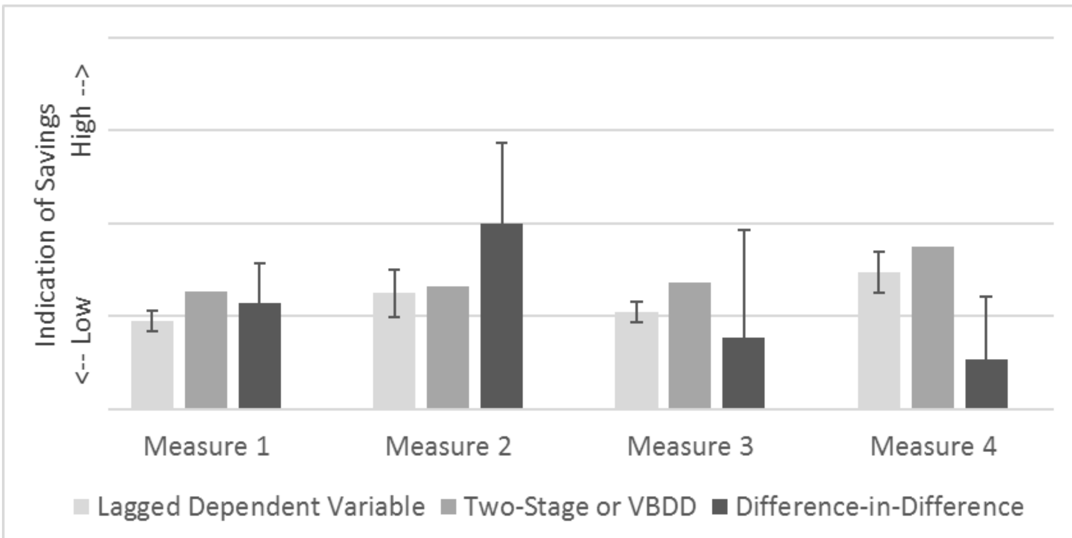

Figure 1: Variation in Savings across Three Model Types and Four Measures
Note: The authors chose not to show the error bars around the two-stage approach, as correctly propagating error through the second regression requires additional considerations that are outside the scope of this work.

Researchers and guidelines can improve this area of variation. Researchers should continue to explore this source of variation, an active area of research, and provide guidance. The UMP should consider to put resources toward making strong recommendations on this topic and should emphasize the potential impact from deviation regarding model selection. Lastly, evaluators should recognize the potential impact on results from model selection.

Comparison Group

Comparison groups account for non-program related changes in consumption, such as those caused by economic changes. Figure 2 provides a conceptual diagram to demonstrate the value of a comparison group. It shows an underlying trend in the data, which would lead to an inaccurate estimate of savings without the control or comparison group data.

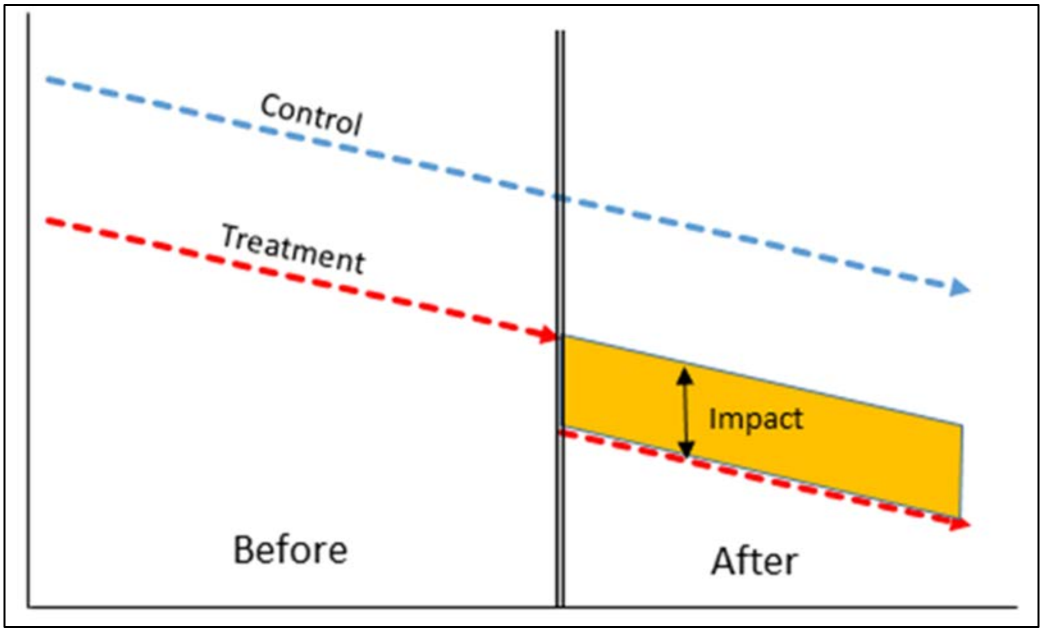

Figure 2: Conceptual Diagram of the Value from a Comparison (or Control) Group

While randomized experimental designs are the ideal approach to account for non-program related changes in energy consumption, such as changes in energy consumption due to economic conditions, that study design is often not feasible (Ho 2007). A randomized assignment would lead to a control group, rather than a comparison group. This paper exclusively discusses comparison groups rather than control groups, because many utility programs do not use randomized studies. In addition, there is more inconsistency around the development of comparison groups than the assignment of control groups.

There is a lot of research underway and progress being made around comparison group development. The authors developed the comparison group for this study using Euclidean distance matching on pre-usage (see Ho et al. 2007 for more discussion on matching) and from the following sources:

- Non-participants, or customers who did not receive rebates or incentives for efficiency upgrades.
- Future participants, or customers who received a rebate or incentive sometime (e.g., a year) after the evaluated participants. When using this approach, it is important to verify the removal of all post-upgrade consumption data for future participants used in the comparison group.

Figure 3 and Figure 4 compare results when using non-participants and future participants to develop comparison groups. Figure 3 shows average daily energy consumption for the months preceding and following measure installation for a treatment and comparison group. This figure shows results for a comparison group derived from non-participants on the left and results for a comparison group derived from future participants on the right. Both graphs in Figure 3 show strong fits before measure installation and clear savings afterward. The similarity of these graphs demonstrates that good matches are achieved using each comparison group data source and that results are similar.

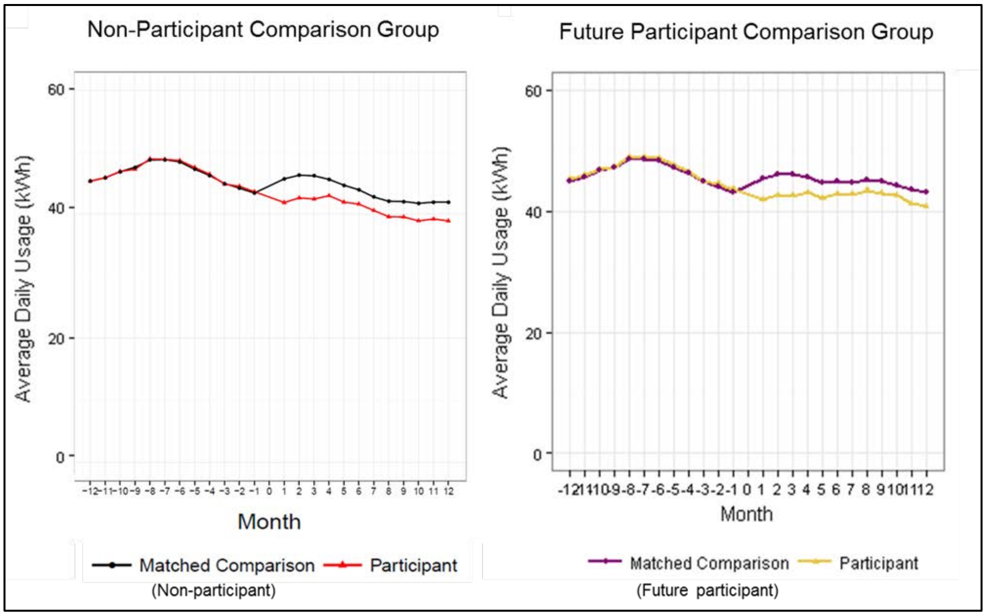

Figure 3: Pre and Post Energy Usage for Participants and Comparison Groups

Because the available non-participant data is only applicable to a subset of the analysis population, Figure 4 shows savings for two different subsets of the population when using matched non-participants and matched future participants as a comparison group. The authors also compare these results to models that do not use a comparison group.

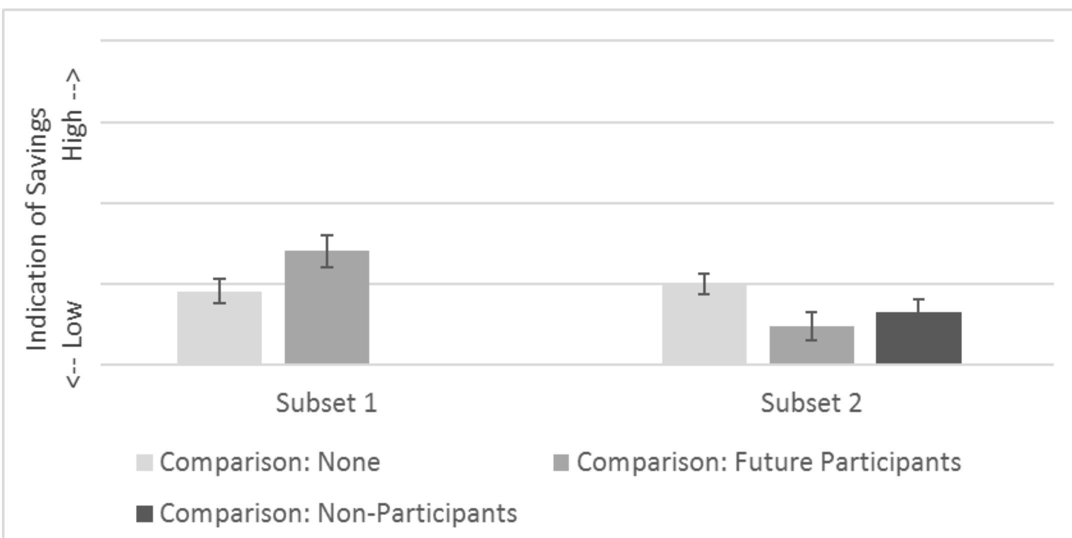

Figure 4: Variation in Savings when Using Different Comparison Group Sources

The results in this study are not statistically different when using different sources to develop a comparison group. However, the results without a comparison group are statistically different. These findings indicate that using a comparison group is important, but the source of the comparison group data may be less important, at least when applying a consistent comparison group matching approach with an adequate sample size.

Based on this result, the authors find that guidelines should require the development of a comparison or control group. Without a comparison group the results cannot be interpreted as savings resulting from a program, but instead as an indication of the change in energy consumption occurring for a given segment of the population at a certain time, some of which may be driven by the program intervention. Additionally, future work should investigate the approach to develop a comparison group, which was not feasible within this study.

Filtering

Filtering approaches are used to avoid findings that are driven by data anomalies. However, the application of data filters and the guidance for data filtering are highly inconsistent (Apex 2016; Cadmus 2012; Cadmus 2015; Greco 2015; Efficiency Valuation Organization 2012; Kelsven 2016; Olig 2016; Patterson 2016; Schiller 2012). UMP Chapter 8 provides a relatively extensive list of filters to apply, and this paper focuses on a few the authors and regional stakeholders believed to be most important and most realistic to investigate. This paper provides results for the following filters by applying each one to the data set, one at a time. These filters are not in use for any other analysis results presented in this paper.

- Pre- to post-consumption change
- Zero consumption bills
- VBDD fit (i.e., whether there is a clear heating signal for each site)
- Pre-period consumption

Figure 5 shows savings when applying filters one at a time across four measures. With a few exceptions, applying each filter does not lead to statistically different results. These results indicate that logically filtering out participants from an analysis data set may not have as substantial an impact on results as model selection.

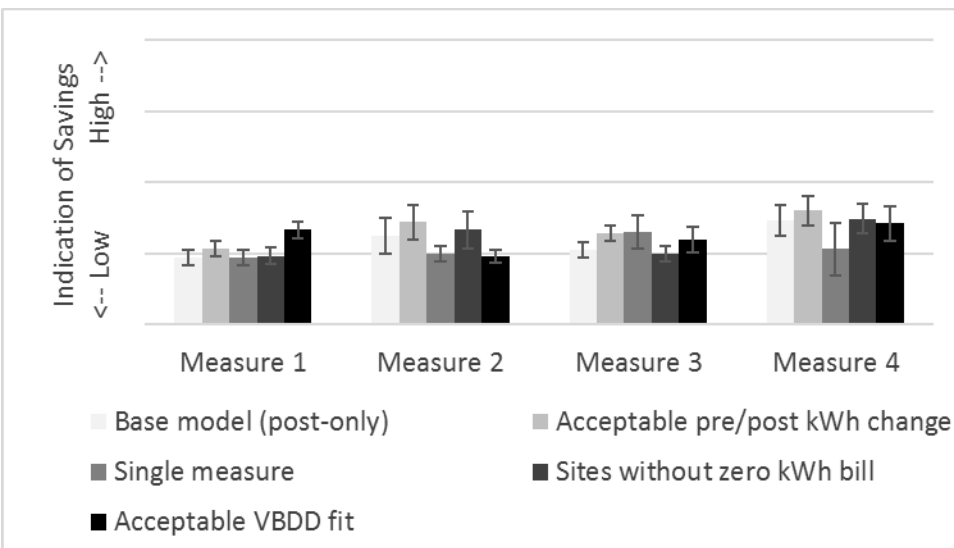

Figure 5: Variation in Savings when Using Select Filters across Measure Groups

Conclusions

The main message from this analysis is that differences in model selection may be driving variability in consumption analysis results and greater consistency could be achieved through targeted research and subsequent updates to existing guidelines. The authors provide more specific conclusions below.

Model types: Model selection has a substantive impact on results. Based on the results of this study, model selection may be a main driver of variation in evaluation results. The authors recommend that model selection and specification receive a high level of attention when developing future evaluation guidelines, that researchers continue to study this issue and that evaluators recognize the possible impact from model selection.

Comparison group: The results here indicate that including a comparison group is likely to have a substantive impact on results and any guidance documents should require a comparison or control group.

Filtering: The results do not indicate that filtering is likely to be a main driver in the variation of savings, at least when the filters are applied in a logical and objective manner.

Additionally, future studies should 1) replicate similar analyses to determine if these findings can be generalized, and 2) investigate the impact from different methods to develop comparison groups.

References

Agnew, K., and M. Goldberg. 2013. *Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol.* Uniform Methods Project, National Renewable Energy Laboratory. https://energy.gov/sites/prod/files/2013/11/f5/53827-8.pdf

Alcott, H., and T. Rodgers. 2012. *The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from* Energy *Conservation.* Cambridge: National Bureau of Economic Research. http://scholar.harvard.edu/files/todd_rogers/files/the_short.pdf

Apex Analytics, LLC. 2016. *Energy Trust of Oregon: Smart Thermostat Pilot Evaluation.* Boulder, CO: Apex Analytics, LLC. http://assets.energytrust.org/api/assets/reports/Smart_Thermostat_Pilot_Evaluation-Final_wSR.pdf

The Cadmus Group, Inc. 2015. *Energy Efficiency in Schools Program: Evaluation, Measurement, and Verification for Duke Energy Kentucky.* Portland: The Cadmus Group, Inc.

The Cadmus Group, Inc. 2012. *Low Income Single Family Program Impact Evaluation: Part of the Massachusetts Residential Retrofit and Low Income Program Area Evaluation.* Portland: The Cadmus Group, Inc. http://ma-eeac.org/wordpress/wp-content/uploads/Low-Income-Single-Family-Program-Impact-Evaluation_Part-of-the-Massachusetts-Residential-Retrofit-Low-Income-Program-Area-Evaluation.pdf

Efficiency Valuation Organization. 2012. *International Performance Measurement and Verification Protocol: Concepts and Options for Determining Energy and Water Savings Volume 1.* Toronto: Efficiency Valuation Organization.
http://www.eeperformance.org/uploads/8/6/5/0/8650231/ipmvp_volume_i__2012.pdf

Gage, L., Baylon, D., Rushton, J., Baker, M., and Spencer, J. 2015. *Cage Match or Happy Couple? Engineering Simulation Models and Billing Analysis.* International Energy Program Evaluation Conference. https://www.iepec.org/wp-content/uploads/2015/papers/018.pdf

Greco, V., and S. Wayland. 2015. *Results for AIC PY6 HPwES Billing Analysis.* Waltham, MA: Opinion Dynamics.

http://ilsagfiles.org/SAG_files/Evaluation_Documents/Ameren/AIU%20Evaluation%20Reports%20EPY6/AIC_PY6_Integrated_Report_FINAL_2016-02-16.pdf

Ho, D., K. Imai, G. King, and E. Stuart. 2007. “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Casual Inference.” *Political Analysis* 15: 199-236. http://gking.harvard.edu/files/matchp.pdf

Kelsven, P., R Weber, and E. Urbatsch. 2016. *Nest Learning Thermostat Pilot Program Savings Assessment.* Bonneville Power Administration. https://www.bpa.gov/EE/Utility/research-archive/Documents/BPA%20-FPUD-Nest-Thermostat-Pilot-Savings-Assessment.pdf

Olig, C., and W. Sierzchula. 2016. *ComEd Home Energy Report Program Evaluation Report.* Chicago: Navigant. http://ilsagfiles.org/SAG_files/Evaluation_Documents/ComEd/ComEd_EPY8_Evaluation_Reports_Final/ComEd_Home_Energy_Report_Opower_PY8_Evaluation_Report_2016-12-22_Final.pdf

Patterson, O., K. Randazzo, and S. Wayland. 2016. *Process and Impact Evaluation of 2014 (PY7) Ameren Illinois Company Behavioral Modification Program.* Waltham, MA: Opinion Dynamics.

Schiller, S. 2012*. Energy Efficiency Program Impact Evaluation Guide: Evaluations, Measurement, and Verification Working Group.* State and Local Energy Efficiency Action Network. https://www4.eere.energy.gov/seeaction/system/files/documents/emv_ee_program_impact_guide_0.pdf

Stewart, J., and A. Todd. 2015. *Chapter 17: Residential Behavior Protocol.* Uniform Methods Project, National Renewable Energy Laboratory. https://energy.gov/sites/prod/files/2015/02/f19/UMPChapter17-residential-behavior.pdf

Violette, D., and P. Rathburn. 2014. *Chapter 23: Estimating Net Savings: Common Practices.* Uniform Methods Project, National Renewable Energy Laboratory. https://energy.gov/sites/prod/files/2015/02/f19/UMPChapter23-estimating-net-savings_0.pdf