The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

January 2012 — September 2016

Michael Li
U.S. Department of Energy
Washington, D.C.

Hossein Haeri and Arlis Reynolds
The Cadmus Group
Portland, Oregon

NREL Technical Monitor: Chuck Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-70472
August 2018

Contract No. DE-AC36-08GO28308

NATIONAL RENEWABLE ENERGY LABORATORY

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

January 2012 — September 2016

Michael Li
U.S. Department of Energy
Washington, D.C.

Hossein Haeri and Arlis Reynolds
The Cadmus Group
Portland, Oregon

NREL Technical Monitor: Chuck Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-70472
August 2018

Contract No. DE-AC36-08GO28308

NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Office of Electricity Delivery and Energy Reliability (OE). The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via www.OSTI.gov.

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Foreword

This report was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home.

Acknowledgments

This report was prepared for the National Renewable Energy Laboratory (NREL) and funded by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy, and the Permitting, Siting and Analysis Division of the Office of Electricity Delivery and Energy Reliability under National Renewable Energy Contract No. DE-AC36-08GO28308. The project was managed by Charles Kurnik of NREL. The Cadmus Group, Inc., managed protocols development with participation from a broad cross-section of experts.

The project engaged a steering committee to provide industry insight and intelligence. We would like to thank all the individuals who participated in the steering committee both past and present, (see list below) for their contributions, and for engaging their organization's staff in the reviews.

Uniform Methods Project Steering Committee

Carmen Best	California Public Utilities Commission
Brian Billing	AEP Ohio
Michael Brandt	Commonwealth Edison
Jessica Burdette	State of Minnesota
Jess Burgess	Consortium for Energy Efficiency
Keith Dennis	National Rural Electric Cooperative Association
Niko Dietsch	Environmental Protection Agency
Linda Ecker	AEP Ohio
Tom Eckman	Regional Technical Forum
Kim Erickson	Consortium for Energy Efficiency
Sandy Fazeli	National Association of State Energy Officials
Hilary Forster	Consortium for Energy Efficiency
Julia Friedman	National Association of State Energy Officials
Carla Frisch	U.S. Department of Energy
Donald Gilligan	National Association of Energy Service Companies
Brian Granahan	Illinois Commerce Commission
Kevin Gunn	Missouri Public Service Commission
David Jacot	Los Angeles Department of Water and Power
Tina Jayaweera	Northwest Power and Conservation Council and Regional Technical Forum
Val Jensen	Commonwealth Edison
Miles Keogh	National Association of Regulatory Utility Commissioners
Steve Kromer	Efficiency Valuation Organization
Jason Kupser	DTE Electric Company
Marty Kushler	American Council for an Energy-Efficient Economy
Michael Li	U.S. Department of Energy
Diana Lin	National Association of State Energy Officials
Julie Michals	E4TheFuture

William Miller	Lawrence Berkeley National Laboratory
William Newbold	DTE Electric Company
Jeffery Orcutt	Illinois Commerce Commission
Mary Ann Ralls	National Rural Electric Cooperative Association
Chuck Rea	MidAmerican Energy Company
Phyllis Reha	Minnesota Public Utilities Commission
Gene Rodrigues	ICF International
Emmett Romine	DTE Electric Company
Steve Rosenstock	Edison Electric Institute
Amy Royden-Bloom	National Association of Clean Air Agencies
Steve Schiller	Lawrence Berkeley National Laboratory
Douglas Scott	Illinois Commerce Commission
Nancy Seidman	Commonwealth of Massachusetts
Rodney Sobin	National Association of State Energy Officials
Cody Taylor	U.S. Department of Energy
Dub Taylor	State of Texas
David Terry	National Association of State Energy Officials
Elizabeth Titus	Northeast Energy Efficiency Partnerships

NREL would also like to thank the industry members that provided feedback on the early drafts of the protocols. Their enduring commitment and the contributions of their staffs greatly enhanced the protocols.

Uniform Methods Project Technical Advisory Group

Mimi Goldberg, DNV GL	Technical Advisory Group
Frank Stern, Navigant	Technical Advisory Group
Terry Fry, Nexant	Technical Advisory Group
M. Sami Khawaja, The Cadmus Group	Technical Advisory Group
Michael Rufo, Itron	Technical Advisory Group
Dick Spellman, GDS Associates, Inc.	Technical Advisory Group
Kevin Warren, Warren Energy	Technical Advisory Group
Tom Eckman, Regional Technical Forum	Net-to-Gross Technical Advisory Group
Steve Schiller, Schiller Consulting	Net-to-Gross Technical Advisory Group
Elizabeth Titus, Northeast Energy Efficiency Partnerships	Net-to-Gross Technical Advisory Group

Special thanks goes out to the following organizations for allowing their staff to dedicate many hours to reading the draft protocols and providing constructive feedback: Commonwealth Edison; Northwest Power and Conservation Council's Regional Technical Forum; National Association of Energy Service Companies; Illinois Commerce Commission; Efficiency Valuation Organization;

Northeast Energy Efficiency Partnerships; Detroit Edison; National Rural Electric Cooperative Association; MidAmerican Energy Company; Southern California Edison; National Association of Clean Air Agencies; Schiller Consulting; Commonwealth of Massachusetts; Alliance to Save Energy; Navigant; BuildingMetrics, Inc.; Itron; Warren Energy; Energy Savvy; Opinion Dynamics; Ridge and Associates; CDP Strategies; Building Metrics; ADM Associates, Inc.; Oracle Utilities; Vermont Energy Investment Corporation; Evergreen Economics; Jacobson Energy Research, LLC; Terra Novum; PECI; ISO New England; Prahl & Associates; Research Into Action, Inc.; TecMarket Works; Natural Resources Defense Council; Pacific Gas and Electric Company; University of Wisconsin; SBW Consulting; Cascade Energy Engineering; Bonneville Power Administration; Strategic Energy Group; and Energy Trust of Oregon.

Finally, NREL also thanks the dozens of individuals and organizations that provided constructive technical feedback during the stakeholder review process. These efforts dramatically improved the quality of the final product.

Suggested Citation

Li, M.; Haeri, H.; Reynolds, A. (2018). *The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-70472. http://www.nrel.gov/docs/fy18osti/70472.pdf

Table of Contents

Chapter 1: Introduction

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016.

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Michael Li

Department of Energy, Office of Energy Efficiency and Renewable Energy

Hossein Haeri and Arlis Reynolds

The Cadmus Group
Portland, Oregon

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68557
September 2017

Contract No. DE-AC36-08GO28308

NATIONAL RENEWABLE ENERGY LABORATORY

Chapter 1: Introduction

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Michael Li
Department of Energy, Office of Energy Efficiency and Renewable Energy

Hossein Haeri and Arlis Reynolds
The Cadmus Group
Portland, Oregon

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68557
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home.

Suggested Citation

Li, Michael; Haeri, Hossein; Reynolds, Arlis. (2017). *Chapter 1: Introduction. The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68557. http://www.nrel.gov/docs/fy17osti/68557.pdf

Acronyms

DOE	U.S. Department of Energy
EM&V	evaluation, measurement, and verification
FEMP	Federal Energy Management Program
IPMVP	International Performance Measurement and Verification Protocol
M&V	measurement and verification
NREL	National Renewable Energy Laboratory
NTG	net-to-gross
SEE Action	State and Local Energy Efficiency Action Network
TRM	technical reference manual
UMP	Uniform Methods Project

Table of Contents

List of Figures

List of Tables

1 Overview

The Uniform Methods Project (UMP) began in 2012 with funding from the U.S. Department of Energy (DOE) to establish a set of model protocols for determining gross energy and demand savings that result from energy efficiency measures and programs implemented through state and utility energy efficiency programs.[1] The protocols provide detailed descriptions of the commonly accepted evaluation methods to help ensure that similar programs are measured in the same way.

The UMP has developed two types of protocols:

1. **Measure-specific protocols** describe recommended evaluation methods for a specific measure, technology, project, or program design type under specified conditions.
2. **Cross-cutting protocols** complement measure-specific protocols by covering evaluation topics, techniques, and technical issues common to all measures.

The methods described in each protocol are—or are among—the most commonly used and accepted in the energy efficiency industry for the specified measure and application conditions.[2] The protocols are authored by experienced evaluators, draw from the existing body of research and best practices for energy efficiency program evaluation, measurement, and verification (EM&V)[3] and are vetted through peer and public stakeholder review processes.

In 2017, the UMP completed a review of all protocols to assess the need for updates or revisions based on changes in the prevailing industry standard methods, lessons learned from recent evaluation activity, and other stakeholder feedback. The UMP revised nine protocols and republished all other protocols to acknowledge their continuity in reflecting the latest evaluation methods.

[1] The UMP protocols are designed primarily for ratepayer-funded energy efficiency programs; however, the protocols can also be used to determine savings from individual projects such as those implemented by energy services companies under a performance contract.
[2] The protocol for data centers is the only exception to this statement. Programs for data centers are relatively new, and the evaluation industry has yet to arrive at a preferred measurement and verification (M&V) approach for the measures offered through these programs. With the data centers protocol, the UMP attempts to describe a preferred approach.
[3] M&V is distinct from evaluation in that it focuses on determining savings for individual measures and projects, while evaluation aims to quantify the impacts of a program.

2 About the Protocols

The methods described in each protocol represent generally accepted practices within the EM&V profession. Although they are not necessarily the *only* manner in which savings can be reliably determined, program administrators, policymakers, and evaluators can adopt these methods with the assurance that they are (1) consistent with accepted practices and (2) have been vetted by experts in the field of energy efficiency program evaluation. If widely adopted, these protocols will help establish a common basis for assessing and comparing the performance and effectiveness of energy efficiency policies and investment decisions across programs, portfolios, and jurisdictions.

These protocols do not provide stipulated values for energy savings. However, their widespread use would provide a common analytic foundation for determining "deemed" values while still allowing for the use of inputs appropriate for the unique circumstances of a project or program.

In general, the measure-specific protocols describe the methods for determining *gross* energy and demand savings. Chapter 21, "Estimating Net Savings: Common Practices," is cross-cutting and describes approaches for determining *net* savings for different measures and programs.

These protocols are designed to provide estimates of gross savings at a high level of rigor; however, they do not prescribe specific criteria for either statistical confidence or precision of savings estimates. Such thresholds are assumed to be set by the stakeholders, as determined by their unique objectives and priorities. Instead, the protocols provide a framework for deciding on and applying such criteria consistently, and for reporting the uncertainty associated with the resulting savings estimates.

3 Rationale

Investment in energy efficiency has increased steadily in the United States over the last decade. In many jurisdictions, energy efficiency now accounts for a significant share of utilities' integrated resource portfolios and, in several jurisdictions, is recognized as the "fuel of first choice," thus amplifying its critical role in electric resource reliability and adequacy.

This trend of increasing investment in energy efficiency will likely continue as utilities strive to meet the energy efficiency resource standards that have been adopted through legislative or regulatory mandates in 26 jurisdictions—and are being considered in several more. In at least half of these jurisdictions, the standards are designed to achieve aggressive savings of 10% or more of forecast load by 2020; in six jurisdictions, savings of more than 20% are expected (ACEEE 2011).

With greater reliance on energy efficiency as a means of meeting future energy demand, there is a growing need for publicly available information on energy efficiency programs, how their savings are determined, and how the achieved savings are reported. Well-documented and consistent use of protocols developed and vetted by experienced practitioners and shared among stakeholders and the public reinforce the reliability of the savings achieved by energy efficiency programs. The UMP protocols offer evaluation methods for determining energy savings based on generally accepted practices in the energy efficiency industry for common measures and programs. Widespread adoption of the UMP protocols also provides a basis for comparing the impacts of energy efficiency portfolios and policy initiatives across the country.

To help reduce the uncertainty associated with determining energy efficiency savings, the UMP protocols also offer guidance for implementing the techniques and interpreting their results.

DOE envisions the following specific goals for this project:

- Offer evaluation methods that strengthen the credibility of energy efficiency program savings calculations.
- Provide clear, accessible, step-by-step procedures to determine savings for the most common energy efficiency measures and programs.
- Support consistency and transparency for how savings are calculated.
- Reduce the costs of developing and managing the evaluation, measurement and verification (EM&V) of energy efficiency projects and programs.
- Allow a comparison of savings across similar programs and measures in different jurisdictions.
- Improve the acceptability of reported energy savings by financial and regulatory communities.

4 The Audiences and Objectives

DOE commissioned the UMP effort to provide, for voluntary adoption, a set of protocols for determining savings that are achieved through state and utility efficiency programs.

By providing a method for evaluating the effectiveness and viability of energy efficiency, these protocols serve stakeholders by:

- Offering regulators a reliable basis and the means for assessing the prudency of rate payer-funded investments in energy efficiency and determining compliance with savings targets.
- Offering utility resource planners and program administrators greater certainty about program performance and reducing planning and regulatory compliance risks.
- Supplying independent EM&V contractors with a standard set of tools and techniques to enhance the accuracy of their findings.
- Providing a learning opportunity for EM&V practitioners and a basis for calculating deemed and algorithm-based savings in technical reference manuals (TRMs) that are being developed or updated in various jurisdictions.
- Providing a resource for program administrators, implementers, and evaluators to determine data collection methods to facilitate the EM&V process.

By making the methods for calculation and verification of savings more transparent and uniform, these protocols can help mitigate the perceived risks of investing in energy efficiency and stimulate greater investment.

5 Definitions

Market participants in the energy efficiency industry (such as end-use energy consumers, project designers, contractors, program implementers and administrators, utility resource planners, and evaluators) may define savings resulting from energy efficiency differently. The UMP uses standard industry definitions consistent with the State and Local Energy Efficiency Action Network (SEE Action) *Energy Efficiency Program Impact Evaluation Guide* to differentiate how savings are reported at the design, implementation, and evaluation stages of a program's life cycle:

- **Projected savings.** Values reported by a program implementer or administrator before the efficiency activities are complete.[4]
- **Gross savings.** Changes in energy consumption that result directly from program-related actions taken by participants in an energy efficiency program, regardless of why they participated.
- **Claimed (gross) savings.** Values reported by a program implementer or administrator after the activities are complete.[5]
- **Evaluated (gross) savings.** Values reported by an independent, third-party evaluator[6] after the efficiency activities and impact evaluation are complete.
- **Net savings.** Change in energy use attributable to a particular energy efficiency program. These changes may implicitly or explicitly include the effects of factors such as free-ridership, participant and nonparticipant spillover, and induced market effects.
- **Net-to-gross (NTG) analysis.** Estimation of the NTG ratio, which is the net savings as a fraction of gross savings.

[4] In certain cases, the projected savings may be based on deemed values approved by regulators.
[5] In certain cases, these savings may have been adjusted by a predetermined NTG ratio.
[6] The designations of "independent" and "third-party" are determined by those entities involved in the use of the evaluations and thus may include evaluators retained by the program administrator or a regulator, for example.

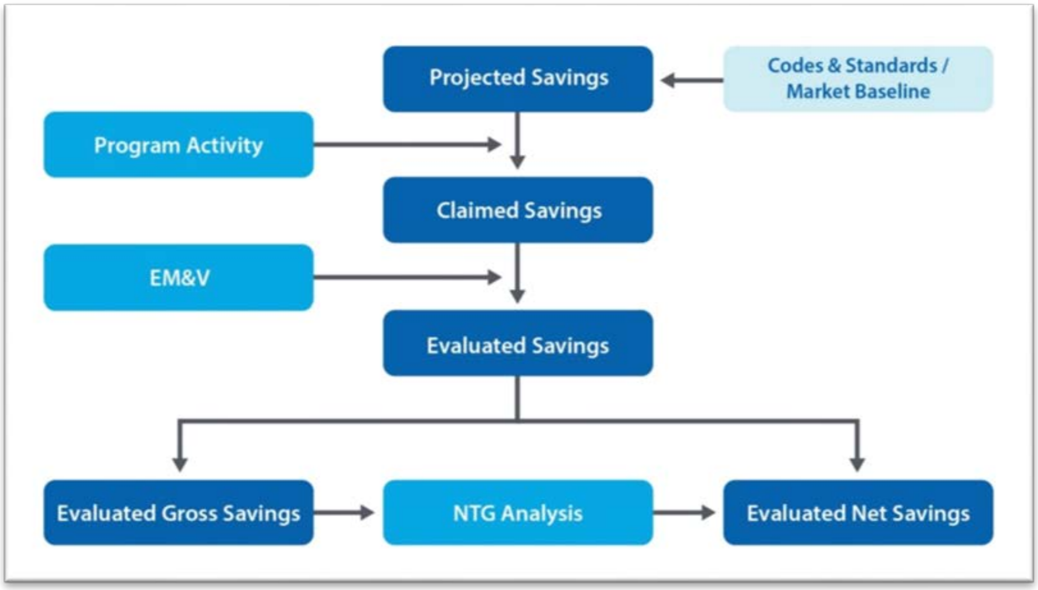


Figure 1. Savings definitions

The UMP protocols focus primarily on estimating evaluated first-year gross savings, except where estimates of net savings may be derived as part of the same method. A more complete discussion of the elements of NTG adjustments and the methods for measuring them are described in *Chapter 21: Estimating Net Savings – Common Practices*. The definition of net savings (for example, whether it includes participant and/or nonparticipant spillover) and the manner in which NTG is applied also vary across jurisdictions as a matter of policy. Therefore, UMP does not offer specific recommendations on how NTG should be measured or applied.

6 Protocol Content

Since its inception in 2012, the UMP created protocols for 17 energy efficiency measures, which are primarily applicable to residential and commercial facilities, and six cross-cutting topics.

6.1 Measure-Specific Protocols

Table 1 shows the 17 measure-specific protocols completed to date. Several of these protocols have been updated to reflect information that has become available since these protocols were first developed. The complete list of protocols developed to date is available on the UMP website.

The UMP prioritized measures that (1) represent a diverse set of end uses in the residential and commercial sectors, (2) are present in most energy efficiency portfolios in nearly all jurisdictions, and (3) have a significant remaining savings potential.

Table 1. UMP Measure-Specific Protocols

Chapter	Protocol Topic	Residential	Commercial	Publish Date
2	Commercial and industrial lighting		X	April 2013; revised 2017
3	Commercial and industrial lighting controls		X	April 2013
4	Small commercial and residential unitary and split system HVAC heating and cooling equipment-efficiency upgrade	X	X	April 2013; revised 2017
5	Residential furnaces and boilers	X		April 2013
6	Residential lighting	X		December 2014; revised 2015; revised 2017
7	Refrigerator recycling	X		April 2013; revised 2017
8	Whole-building Retrofit with consumption data analysis	X	X	April 2013; revised 2017
14	Chillers		X	September 2014
15	Commercial new construction		X	September 2014
16	Retrocommissioning		X	September 2014
17	Residential behavior	X		January 2015; revised 2017
18	Variable frequency drive		X	November 2014
19	HVAC controls (DDC/EMS/BAS)		X	November 2014
20	Data center IT efficiency measures		X	January 2015

Chapter	Protocol Topic	Residential	Commercial	Publish Date
22	Compressed air		X	November 2014; revised 2017
23	Combined heat and power		X	November 2016
24	Strategic energy management		X	May 2017

6.1.1 Protocol Organization

Each UMP protocol explains the underlying technology, the end uses affected by the measure, the method for calculating the measure's gross savings, and the data requirements. Also, each protocol attempts to provide sufficient detail without being overly prescriptive, allowing flexibility and room for professional judgment.

The content in the measure-specific protocols is organized in a similar structure to provide consistency. Each protocol provides the following information:

- **Measure description.** A brief description of the measure covered by the protocol
- **Application conditions of protocol.** Details on the types of delivery channels, program structures, or other conditions that are or are not covered by the protocol
- **Savings calculations.** The prevailing algorithm(s) used to estimate energy savings with an explanation of the parameters
- **M&V plan.** The recommended evaluation approach, including the International Performance Measurement and Verification Protocol (IPMVP) Option, where appropriate, for determining values for the parameters required in the savings calculation
- **Sample design.** Overview of considerations on how to segment the population to provide a representative sample for evaluation, which is discussed in conjunction with the M&V Plan in some protocols
- **Other evaluation issues.** Any additional information deemed pertinent by the author(s) and/or reviewers, including brief discussions of persistence or NTG considerations; often this information is supplemented by the crosscutting protocols
- **References.** Complete citations of reference and resource materials discussed in the protocols, including example evaluation reports that demonstrate the recommended evaluation method.

In addition, the protocols revised in 2017 include two new sections:

- **Revisions.** Summary of key changes from the previous version of the protocol
- **Looking Forward.** Discussion of upcoming or potential changes based on ongoing research, new evaluation tools, future changes in the market, or other experimental methods.

Each measure is unique; therefore, some protocols have additional sections to provide more details on specific areas of interest or consideration.

6.2 Cross-Cutting Protocols

Cross-cutting protocols outlined in Table 2 complement the measure-specific protocols by covering technical issues and topics common to all measures. These crosscutting topics provide guidance on specific topics as stand-alone documents or may be referenced in measure-specific protocols. These supplemental, crosscutting discussions help extend the measure-specific method for determination of savings to evaluating whole programs.

Table 2. UMP Cross-Cutting Protocols

Chapter	Protocol Topic	Publish Date
1	Introduction	April 2013; revised 2017
9	Metering	April 2013
10	Calculation of peak demand and time impacts	April 2013; revised 2017
11	Sample design	April 2013
12	Survey design and implementation	April 2013
13	Assessing persistence and other evaluation issues	April 2013
21	Common practices in estimating net savings	September 2014; revised 2017

6.3 Relationship to Other Protocols

The UMP protocols are based on long-standing EM&V practices and well-established scientific principles. They draw from and build on previous attempts to develop systematic approaches to estimating the impacts of energy efficiency. Those efforts were conducted by various entities, including Oak Ridge National Laboratory (ORNL 1991), the Electric Power Research Institute (EPRI 1991), the U.S. Environmental Protection Agency (EPA 1995), and DOE (1996, 2008).

Several of these protocols were developed to address specific policy objectives, such as the verification of utility program savings, the determination of savings from special performance contracts, and environmental compliance.

The UMP protocols also draw on the IPMVP (DOE 2002). Each measure-specific protocol identifies the IPMVP Option with which it is associated, expands on the IPMVP Option by adding measure-specific detail, and describes the procedures for application to measure-, program- and portfolio-level evaluations.

In addition, the UMP protocols draw from and build on EM&V protocols developed to establish standards and consistency for evaluation activities within specific jurisdictions, These jurisdictional protocols include resources developed in California, by the Regional Technical Forum in the Pacific Northwest, and by the Northeast Energy Efficiency Partnerships' EM&V Forum.

A valuable companion document to the UMP protocols is the *SEE Action Energy Efficiency Program Impact Evaluation Guide* (SEE Action 2012). The SEE Action guide provides both an introduction to and a summary of the practices, planning, and associated issues of documenting energy savings, demand savings, avoided emissions, and other non-energy benefits resulting from end-use energy efficiency programs.

Designed to be complementary with the *SEE Action Energy Efficiency Program Impact Evaluation Guide*, the UMP protocols are more detailed and specific for particular measures and projects. (The preparation of these protocols was closely coordinated with that guide.)

For many technologies, evaluation tools and methods continue to improve, and the industry will continue to benefit from advancements to evaluation methods so system performance can be estimated more accurately in the future. The evaluation methods will continue to evolve in response to these changes.

7 About Evaluation, Measurement, and Verification Budgets

The EM&V effort—and expenditures—should be scaled to both the program being evaluated and the accuracy necessary to inform the decision for which evaluation results matter. The value of the information provided by the EM&V activity is determined by the resource benefits of the program and the particular policy objectives and research questions the EM&V activity aims to address.

Historically, the costs of determining energy savings are embedded in the larger context of evaluation activities undertaken as part of large-scale program portfolios. The level of effort and the corresponding cost of implementing the UMP protocols vary. In addition, EM&V costs vary depending on the regulatory requirements that dictate the levels of statistical confidence and precision. A survey of evaluation budgets for large program portfolios available from regulatory filings in several jurisdictions indicates portfolio-level EM&V expenditures ranging from 2% of total portfolio costs in Indiana to 6% of total portfolio costs in other jurisdictions.[7]

These budget estimates should be considered as only rough guidance as they are mostly self-reported and the definitions of cost elements may vary. This is particularly true considering how internal verification processes may differ from independent, third-party evaluations (SEE Action 2012, Section 7.5.2).

Evaluation resource requirements also depend on how often evaluations are conducted. The frequency evaluations are performed depends on a number of considerations, including the type and complexity of the measure and its expected contribution to portfolio savings, the uncertainty about the savings, the lifecycle stage of the program in question, and regulatory requirements. UMP has no specific recommendation about how often programs should be evaluated.

[7] Similar estimates are also available for Illinois (3%), Indiana (5%), Michigan (5%), and Pennsylvania (2%-5%), and Arkansas (2%-6%).

8 Considering Resource Constraints

The UMP protocols draw on best practices to recommend approaches for providing accurate and reliable estimates of energy efficiency savings, within the confines of the typical evaluation budget for that particular program. However, the UMP protocols do not offer recommendations about the levels of rigor and the specific criteria for accuracy of the savings estimates. Those issues are largely matters of policy, ease and cost of data acquisition, and availability of resources.

To provide maximum flexibility, protocols may contain recommendations for alternative, lower-cost means of deploying the protocol, such as relying on secondary sources of data for certain parameters and identifying guidelines for selecting appropriate sources of such data. Practitioners should document when they have used these alternative means.

The costs of deploying the UMP protocols vary depending on the features of the energy efficiency program being evaluated, the participant characteristics, the desired levels of rigor and accuracy, and whether the evaluator employed any alternatives. Thus, cost estimates for implementing the protocols are not provided. Instead, the utilities and program administrators adopting the protocols should consider benchmarking their programs and gauging their EM&V budgets against those of other entities with experience in conducting EM&V for similar programs.

8.1 Options for Small Program Administrators

UMP recognizes that even the lower-cost options provided in the UMP protocols may be impractical where resources are constrained or programs are small (such as those offered by small utilities) (GDS Associates, Inc. 2012).[8] In these circumstances, program administrators may consider using deemed savings values from:

- TRMs created by regional or state entities
- Evaluations of similar programs performed by other regional utilities. (These can serve as the basis for determining energy efficiency savings, provided that the evaluation still verifies the installation and proper operation of the energy efficiency measure or device.)

Deemed savings may be adjusted to allow for climate or other factors (regional or economic/demographic) that differ from one jurisdiction to another. Given the differences in how TRMs determine savings for identical measures, program administrators should use deemed savings values based on calculations and stipulated values derived using the UMP protocols when possible. Those using this approach should update their deemed savings values periodically to incorporate changes in appliance and building codes and the results of new EM&V studies (such as the primary protocols developed under the UMP or other secondary sources).

Where possible, program administrators may consider other cost-saving measures, such as pooling EM&V resources and jointly conducting evaluations of similar programs through local

[8] According to the Small Business Administration, small utilities are currently defined as electric-load-serving entities with annual sales of less than 4 million megawatt-hours.

associations. This resource-pooling has been done successfully with small utilities in California, Minnesota, Michigan, and the Pacific Northwest, as well as across the Northeast region via the Northeast Energy Efficiency Partnerships.[9]

Small or resource-constrained utilities and program administrators may also consider either coordinating with larger, regional utilities or adopting the results of evaluations of similar programs implemented by larger utilities.

[9] http://www.neep.org/

9 Project Management and Oversight

The UMP is funded by DOE and is being managed by the National Renewable Energy Laboratory (NREL). Since the project launched in 2011, the Cadmus Group, Inc., has managed technical aspects of the project, including protocol development, and provides technical oversight. The management structure was designed to be inclusive of a broad set of stakeholders to engage expertise and input across the industry and ensure technical excellence.

Figure 2 describes the management structure for the UMP:

- NREL manages membership and communication with the project Steering Committee and administers the public comment process.
- Cadmus manages the subject matter technical experts who develop protocols and the project Technical Advisory Group.
- The Steering Committee[10] is made up of thought leaders with perspectives on policy issues who approve project structure, guide selection of measures or topics for protocols, review final work products, and promote protocol adoption.
- The Technical Advisory Group[11] reviews all protocols and provides EM&V guidance on the validity, usability, and attribution components through the development process for each protocol.

As project sponsor, DOE oversees all aspects of the project, articulates overarching project goals, and ensures the UMP products meet DOE policy objectives.

Figure 2. UMP Management Structure

[10] Members of the Steering Committee are listed on the UMP website: https://www.nrel.gov/ump/steering-committee.html

[11] Members of the Technical Advisory Group are listed on the UMP website: https://www.nrel.gov/ump/technical-advisory.html

The following sections describe the management strategies employed to facilitate the final appeal and acceptance of the UMP work products.

9.1 Project Oversight by Variety of Stakeholders

NREL formed the project steering committee to provide general direction and guidance. The steering committee consists of regulators, utility managers, energy planners and policymakers, and representatives of industry associations. Members of the UMP Steering Committee are listed on the UMP website: https://www.nrel.gov/ump/steering-committee.html.

9.2 Authorship by Experts

Nationally recognized experts on specific energy efficiency measures, technologies, and research techniques draft each protocol in consultation with their peers. Each protocol represents the best method as agreed to by several leading experts, not just the lead author.

9.3 Review by Technical Advisory Group

A technical advisory group made up of experts from major consulting firms engaging in EM&V throughout North America reviews draft and final protocols to verify that the proposed method is a valid way to measure savings, and to ensure the protocol is written in a way that is understandable to evaluators that will use it.

Members of the UMP Technical Advisory Group are listed on the UMP website: https://www.nrel.gov/ump/technical-advisory.html.

9.4 Review by Stakeholders

All protocols are subject to a public review process, administered by NREL, which allows stakeholders to provide feedback on draft protocols before they are released in their final form.

9.5 Monitoring Use and Adoption

To monitor protocol use and adoption, Cadmus tracks references to protocols in various program and evaluation materials, including frameworks and guidelines, EM&V requests for proposals, EM&V workplans and reports, and TRMs, as well as other citations in industry reports or articles. The project maintains a record of such adoptions and periodically reports on known uses.

9.6 Protocol Refresh and Revision

To ensure the project protocols remain useful and up-to-date, especially as evaluation methods evolve to employ new tools and techniques, NREL maintains an email to receive feedback from stakeholders, periodically soliciting feedback from the Steering Committee, Technical Advisory Group, protocol experts, and industry.

In 2016, three years after the first set of protocols was published, the project team polled past UMP authors and other contributors to collect feedback on the need for revisions to existing protocols. Based on feedback from the authors and other stakeholders, the project initiated a process in 2017 to:

- Revised nine existing protocols to incorporate changes and clarifications in the methods, relevant new research and updated references, and other stakeholder feedback; and
- Republish the remaining protocols to acknowledge their continued viability of those protocols as originally published.

The project continues to take feedback through the project website (https://energy.gov/eere/about-us/ump-home) and email address (ump@nrel.gov).

References

ACEEE (2011). *Energy Efficiency Resource Standards: A Progress Report on State Experience*. Report Number U112. Washington, D.C.: American Council for an Energy Efficiency Economy. http://www.aceee.org/research-report/u112.

Birr, D.; Donahue, P. *Meeting the Challenge: How Energy Performance Contracting Can Help Schools Provide Comfortable, Healthy, and Productive Learning Environments*. The National Association of Energy Services Companies, Washington, D.C.

DOE (1996). *The North American Energy M&V Protocols*. Washington, D.C.: U.S. Department of Energy, DOE-GO 10096-248.

DOE (2002). *International Performance Measurement and Verification Protocols, Concepts and Options for Determining Water and Energy Savings, Vol. 1*. Washington, D.C.: U.S. Department of Energy. http://www.nrel.gov/docs/fy02osti/31505.pdf

DOE (2008). *Federal Energy Management Program (FEMP) M&V Guidelines: Measurement and Verification for Federal Energy Projects Version 3.0*, Washington, D.C.: U.S. Department of Energy.

SEE Action (2011). National Energy Efficiency Evaluation, Measurement and Verification (EM&V) Standard: Scoping Study of Issues and Implementation Requirements at https://www4.eere.energy.gov/seeaction/system/files/documents/emvstandard_scopingstudy.pdf

EPA (1995). *Conservation Verification Protocols, Version 2*, Washington, D.C.: U.S. Environmental Protection Agency, EPA-430/B-95-012.

EPRI (1991). *Impact Evaluation of Demand-Side Management Programs, Vol. 1: A Guide to Current Practice*. Palo Alto, CA: Electric Power Research Institute, EPRI CU-7179.

GDS Associates, Inc. (2012). *GDS Analysis of Proposed Department of Energy Evaluation, Measurement and Verification Protocols.* Marietta, GA: GDS Associates, Inc. https://www.nreca.coop/wp-content/uploads/2013/12/EMVReportAugust2012.pdf

ORNL (1991). *Handbook of Evaluation of Utility DSM Programs*. Oak Ridge, TN: Oak Ridge National Laboratory, ORNL/CON-336.

SEE Action (2012). *Energy Efficiency Program Impact Evaluation Guide*. Prepared by Evaluation, Measurement, and Verification Working Group. https://www4.eere.energy.gov/seeaction/system/files/documents/emv_ee_program_impact_guide_0.pdf

Violette, D. M.; Rathbun, P. (2014). Chapter 17: Estimating Net Savings: Common Practices. Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures; Period of Performance: September 2011-December 2014. Golden, CO: National Renewable Energy Laboratory, NREL/SR-7A40-62678. http://www.nrel.gov/docs/fy14osti/62678.pdf

Chapter 2: Commercial and Industrial Lighting Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Dakers Gowans
Left Fork Energy
Harrison, New York

Chad Telarico
DNV GL
Mahwah, New Jersey

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68558
October 2017

Contract No. DE-AC36-08GO28308

Chapter 2: Commercial and Industrial Lighting Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Dakers Gowans
Left Fork Energy
Harrison, New York

Chad Telarico
DNV GL
Mahwah, New Jersey

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68558
October 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy
and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter authors wish to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Jeff Cropp and M. Sami Khawaja of Cadmus
- Miriam Goldberg of DNV GL
- Dick Spellman of GDS
- Frank Stern of Navigant
- Kevin Warren of Warren Energy.

Suggested Citation

Gowans, D.; Telarico, C. (2017). *Chapter 2: Commercial and Industrial Lighting Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68558. http://www.nrel.gov/docs/fy17osti/68558.pdf

Acronyms

CF	coincidence factor
COP	coefficient of performance
CT	current transformer
DOE	U.S. Department of Energy
HGSF	heat gain space fraction
HID	high-intensity discharge
HOU	hours of use
HVAC	heating, ventilating, and air conditioning
ISR	in-service rate
kW	kilowatt
kWh	kilowatt-hour
LED	light-emitting diode
LPD	lighting power density
M&V	measurement and verification
PCF	peak coincidence factor
UMP	Uniform Methods Project
W	watt

Protocol Updates

The original version of this protocol was published in April 2013.

This chapter has been updated to incorporate the following revisions:

- Added new section for midstream programs.
- Provided additional detail on the recommended duration of metering.
- Changed the recommended minimum metering time from two weeks to four weeks.
- Added an alternative approach to estimating interactive effects through the use of an engineering equation (in addition to the current approach that uses stipulated factors).
- Provided guidance for creating fixture codes for light-emitting diode fixtures not currently found in most look-up tables.
- Updated the protocol on reporting uncertainty based on new material from the International Performance Measurement and Verification Protocol.
- Applied the controls requirements of International Energy Conservation Code 2012 / 90.1-2010 to estimate of baseline hours of use for new construction projects.

Table of Contents

List of Figures

List of Tables

1 Measure Description

The Commercial and Industrial Lighting Evaluation Protocol (the protocol) describes methods to account for gross energy savings resulting from the programmatic installation of efficient lighting equipment in large populations of commercial, industrial, and other nonresidential facilities. This protocol does not address savings resulting from changes in codes and standards, or from education and training activities. A separate Uniform Methods Project (UMP) protocol, *Chapter 3: Commercial and Industrial Lighting Controls Evaluation Protocol*, addresses methods for evaluating savings resulting from lighting control measures such as adding time clocks, tuning energy management system commands, and adding occupancy sensors.

Historically, lighting equipment has accounted for a significant portion of cost-effective, electric energy efficiency resources in the United States, a trend likely to continue as old technologies improve and new ones emerge. By following the methods presented here, the energy savings from lighting efficiency programs in different jurisdictions or regions can be measured uniformly, providing planners, policymakers, regulators, and others with sound, comparable data for comprehensive energy planning. Also, the methods here can be scaled to match the evaluation costs to the value of the resulting information.[1]

An energy efficiency measure is defined as a set of actions and equipment changes that result in reduced energy use—compared to standard or existing practices—while maintaining the same or improved service levels for customers or processes. Energy-efficient lighting measures in existing facilities deliver the light levels (illuminance and spatial distribution) required for activities or processes at reduced energy use, compared to original or baseline conditions. In new construction, “original or baseline condition” usually refers to the building codes and standards in place at the time of construction.

Examples of energy-efficient lighting measures in commercial, industrial, and other nonresidential facilities include:

- Retrofitting existing, linear, fluorescent fixtures with efficacious[2] lamps and ballasts, or delamping over-lit spaces
- Replacing compact fluorescent lamps with screw-in light-emitting diodes (LED) lamps
- Replacing metal halide high-bay fixtures with efficacious LED high-bay equipment.

In practice, lighting retrofit projects and new construction projects commonly implement lighting fixture and lighting controls measures concurrently. This protocol accommodates these mixed measures.

[1] As discussed in the “Considering Resource Constraints” section of the UMP *Chapter 1: Introduction*, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

[2] Efficiency of lighting equipment is expressed as “efficacy,” in units of lumens per Watt, where lumens are a measure of light output.

2 Application Conditions of the Protocol

Energy efficiency lighting programs result in the installation of commercial, industrial, and nonresidential high-efficiency lighting measures in customer facilities. The programs can take advantage of varying delivery mechanisms, depending on target markets and customer types. Primarily, these mechanisms can be distinguished by the parties receiving incentive payments from a program. Although the methods described in this protocol apply to all delivery mechanisms, issues verifying customer and baseline equipment data vary.

2.1 Common Program Types

The following are descriptions of common program types used to acquire lighting energy and demand savings.

2.1.1 Incentive and Rebate

Under this model, implementers pay program participants in target markets to install lighting measures. This type of program is generally referred to as a downstream program. A participant receives either an incentive payment, based on savings ($/kilowatt-hour [kWh]), or a rebate for each fixture or lamp ($/fixture, $/lamp). The terms incentive and rebate sometimes are used interchangeably, but generally, incentives are calculated based on project savings and rebates are based on equipment installed. Examples of participants include contractors, building owners, and property managers.

Savings can be estimated using simple engineering calculations. Some programs include a measurement and verification (M&V) process, in which key parameters—such as hours of use (HOU), baseline, and retrofit fixture wattages—are verified or measured, or both, as part of project implementation.

Rebate programs typically pay for specific lighting equipment types (for example, a 4-foot, four-lamp, T5 electronic ballast fixture), often after they have been installed, so assumptions must be made about baseline or replaced equipment. The result is a tradeoff: increased administrative efficiency for less certainty about baseline conditions (and therefore, savings).

Incentive programs often collect more detailed baseline data than do rebate programs. Typically, these data include baseline and retrofit equipment wattages and HOUs, which facilitate determination of savings impacts.

Although rebate programs typically track useful information about replacement lighting equipment, they may not collect baseline data.

2.1.2 Upstream Buy-Down

In upstream buy-down scenarios, programs pay incentive dollars to one or more entities such as retail outlets, distributors, or manufacturers in the lighting equipment market distribution chain. The upstream approach has been widely used in the residential sector, particularly for compact fluorescent lamp (CFL) commercial and industrial lighting programs.

Upstream programs do not interact with the end-use customers purchasing energy-efficient equipment, making the determination of baseline conditions and installation rates more difficult than for incentive and rebate programs. Program planners, implementers, and evaluators estimate these parameters using regional bulk sales data, market research studies, assessments of current product standards and practices, and experience with other programs.

A subset of upstream programs is the midstream model where incentives are paid to distributors for sales of pre-approved qualified products. Purchasers are contractors and commercial, institutional and governmental accounts. Programs can leverage the relationship between distributor, purchaser, and end-user to require information about equipment sales and the end-use installation. Midstream lighting programs are increasingly included in ratepayer-funded energy efficiency portfolios.

2.1.3 Direct Install

Under this delivery approach, contractors, acting on a program's behalf, install energy-efficient lighting equipment in customer facilities. The programs pay contractors directly. Customers receive a lighting retrofit at reduced cost. Direct-install programs often target hard-to-reach customers—typically small businesses—that are overlooked by contractors working with incentive and rebate programs.

Direct-install programs can usually collect precise information about baseline and replacement equipment, and the program implementers may have reasonable estimates of annual operating hours. Data, when collected, can be used directly by impact evaluation researchers.

2.2 Program Target Markets

In addition to being distinguished by their delivery mechanisms, commercial, industrial, and non-residential lighting programs can be classified by targeting retrofits (serving existing facilities) and new construction markets. Program delivery types described above apply to retrofit programs. New construction programs also employ incentives and rebates (and customers may benefit from upstream buy-downs) to improve lighting energy efficiency.

New construction programs present evaluators with a dilemma in establishing baselines for buildings that have yet to be built. The problem is addressed by referring to new construction energy codes for commercial, industrial, and nonresidential facilities (usually by referencing International Energy Conservation Code or ASHRAE Standard 90.1). The codes define lighting efficiency, primarily in terms of lighting power density (lighting watts/ft^2), calculated using simple spreadsheets. Other federal, state, and local standards may set additional baseline constraints on lamps, ballasts, and fixture efficiency/efficacy.

3 Savings Calculations

Project and program savings for lighting and other technologies result from the difference between the energy consumption that would have occurred had the measure not been implemented (the baseline) and the consumption occurring after the retrofit. Energy calculations use the following fundamental equation:

Equation 1. Energy Savings = (Baseline-Period Energy Use – Reporting-Period Energy Use) ± Adjustments

The equation's adjustment term calibrates baseline or reporting use and demand to the same set of conditions. Common adjustments account for changes in schedules, occupancy rates, weather, or other parameters that can change between baseline and reporting periods. Adjustments commonly apply to heating, ventilating, and air conditioning (HVAC) measures, but less commonly to lighting measures, or are inherent in algorithms for calculating savings.

Regulators and program administrators may require that lighting energy efficiency programs report demand *and* energy savings. Demand calculations use the following fundamental equation:

Equation 2. Demand Savings = (Baseline-Period Demand – Reporting-Period Demand) ± Adjustments

Demand savings, which is calculated for one or more time-of-use periods, is typically reported for the peak period of the utility system serving the efficiency program customers.

3.1 Algorithms

The following equations calculate first-year energy and demand on-site savings for lighting measures in commercial, industrial, nonresidential facilities:

3.1.1 Energy Savings

Equations in this section are used to calculate first-year energy savings for lighting measures.

Equation 3. Lighting Electric Energy Savings

$$kWh\ Save_{light} = \left[\sum_{u,i}\left(\frac{fix\ watt_{base,i}\cdot qty_{base,i}}{1000}\cdot HOU_{base}\right)_u - \sum_{u,i}\left(\frac{fix\ watt_{energy\ efficient,i}\cdot qty_{ee,i}}{1000}\cdot HOU_{ee}\right)_u\right]\cdot ISR$$

Where:

$kWh\ Save_{light}$ = Annual kWh savings resulting from the lighting efficiency project

$fix\ watt_{base,\ energy\ efficient,\ i}$ = Fixture wattage, baseline or energy-efficient, fixture type i

$qty_{base, energy efficient, i}$ = Fixture quantity, baseline or energy-efficient, fixture type i

u = Usage group, a collection of fixtures sharing the same operating hours and schedules, for example all fixtures in office spaces or hallways

$HOU_{base, energy efficient}$ = Annual hours of use, baseline or energy-efficient, usually assumed unchanged from baseline unless new controls are installed

ISR = In-service rate, the percentage of incentivized lamps or fixtures that are installed and operating. Applies to upstream buy-down programs, normally not applicable for incentive and rebate programs

Equation 4. Interactive Cooling Energy Savings for Interior Lighting

$$kWh\,Save_{interact-cool} = \frac{HGSF \cdot kW\ Save_{light} \cdot HOU_{cool}}{COP_{cool}}$$

OR

$$kWh\,Save_{interact-cool} = kWh\,Save_{light} \cdot IF_{kWh,cool}$$

Equation 5. Interactive Heating Energy Penalty for Interior Lighting

$$kWh\,Save_{interact-heat} = \frac{-HGSF \cdot kW\ Save_{light} \cdot HOU_{heat}}{COP_{heat}}$$

OR

$$kWh\,Save_{interact-heat} = kWh\,Save_{light} \cdot IF_{kWh,heat}$$

Where:

kWh $Save_{interact\text{-}cool}$ = Interactive cooling energy impact due to a lighting efficiency project

kWh $Save_{interact\text{-}heat}$ = Interactive heating energy impact from a lighting efficiency project

kW $Save_{light}$ = Connected kW savings ($kW_{base} - kW_{efficient}$) due to a lighting efficiency project

HOU_{cool} = Hours of use of lighting equipment coincident with cooling system operation

HOU_{heat} = Hours of use of lighting equipment coincident with heating system operation

$IF_{kWh, cool}$ = Interactive cooling factor: the ratio of cooling energy reduction per unit of lighting energy reduction

$IF_{kWh,\ heat}$ = Interactive heating factor: the ratio of heating energy increase per unit of lighting energy

COP_{cool} = Cooling system coefficient of performance

COP_{heat} = Heating system coefficient of performance

HGSF = Heat gain space fraction: the percent of lighting wattage that is transferred to the conditioned space as thermal energy.

The protocol provides two options each for calculating cooling and heating interactive effects: an engineering approach and a stipulated factor approach.

In the engineering approach, the HGSF represents the percentage of the lighting energy that is thermal energy added to the conditioned space. According to a 2007 ASHRAE study (Chantrasrisalai and Fisher 2007), the percentage of lighting energy transmitted to the space can range from 12% to 100% depending on the type of light fixture at typical operating conditions.[3] The protocol recommends a default of 70% to 80% HGSF. Calculating building-specific HGSF values is unusual due to the level of effort required.

Interactive effects apply only to interior lighting that operates in mechanically cooled or heated spaces.

Equation 6. Total Annual Energy Savings Due to Lighting Project

$$kWh\ Save_{total=}kWh\ Save_{light} + kWh\ Save_{interact-cool} + kWh\ Save_{interact-heat}$$

3.1.2 Electric Peak Demand Savings

The equations in this section are used to calculate first-year electric peak demand savings for lighting measures. Additional information is available in the Uniform Methods Project (UMP) *Chapter 10: Peak Demand and Time-Differentiated Energy Savings Cross-Cutting Protocol*.

Equation 7. Lighting Electric Peak Demand Savings

$$kW\ Peak\ Save_{light} = CF \cdot \sum_{u,i} \left(\frac{fix\ watt_{base,i} \cdot qty_{base,i}}{1000} - \frac{fix\ watt_{ee,i} \cdot qty_{ee,i}}{1000} \right)_u \cdot ISR$$

Where:

CF = coincidence factor, the fraction (0.0 to 1.0) of connected lighting load turned on during a utility peak period

[3] A value less than 100% means that a portion of the lighting energy is being transferred into the plenum rather than the conditioned space.

Equation 8. Interactive Electric Cooling Demand Savings for Interior Lighting

$$kW\ Peak\ Save_{interact-cool} = \frac{HGSF \cdot kW\ Peak\ Save_{light}}{COP_{cool}}$$

OR

$$kW\ Peak\ Save_{interact-cool} = kW\ Save_{light} \cdot IF_{kW,cool}$$

Where:

> $kW\ Peak\ Save_{interact\text{-}cool}$ = Interactive electric cooling peak demand impact from a lighting efficiency project
>
> HGSF = Heat Gain Space Fraction: the percent of lighting wattage that is transferred to the conditioned space as thermal energy
>
> COP_{cool} = Cooling system coefficient of performance
>
> $IF_{kW, cool}$ = Interactive cooling factor, ratio of cooling demand reduction per unit of lighting demand reduction during the peak period resulting from the reduction in lighting waste heat removed by an HVAC system
>
> Interactive effects apply only to interior lighting operating in mechanically cooled spaces. Interactive heating effects are often ignored in North America because heating equipment is typically nonelectric and heating demand may not coincide with utility system peaks.

Equation 9. Total Electric Peak Demand Savings Due to Lighting Project

$$kW\ Peak\ Save_{total=}kW\ Peak\ Save_{light} + kW\ Peak\ Save_{interact-cool}$$

4 Role of the Lighting Program Implementer

Successful application of this protocol requires collecting standard data in a prescribed format as part of the implementation process. The protocol further requires tracking project and program savings estimated on the basis of those standard data.

The implementer is responsible for ensuring necessary data are collected to track program activity and to calculate savings at the project level. The implementer is also responsible for maintaining a program activity record, including anticipated savings by project.

4.1 Program Implementer Data Requirements

The protocol recommends the program implementer collect and archive, for all projects, all data needed to execute the savings algorithms. These data are:

- Baseline fixture inventory, including fixture wattage
- Baseline fixture quantities
- Baseline lighting HOU
- Efficient fixture inventory, including wattage
- Efficient fixture quantities
- Efficient lighting HOU
- Usage group assignments
- Heating and cooling equipment types
- Interactive factor for cooling, or cooling equipment COP and lighting HOU coincident with cooling equipment operation (optional)
- Interactive factor for heating, or heating equipment COP and lighting HOU coincident with heating equipment operation (optional).

Facilities—or spaces within facilities where the project is installed—are classified as cooled/uncooled or heated/unheated. Recording information about heating and cooling equipment and fuel types for each facility or space allows for more precise estimation of interactive effects. Implementers may elect to use a program-level default values for the percent of space that is heated or cooled. The values can be based on earlier studies or evaluation reports for similar populations.

Note that some of the information will not be available for some program types (e.g., baseline fixture information for new construction, upstream, or midstream programs). See Section 8 for recommendations for midstream programs.

4.2 Implementation Data Collection Method

The protocol recommends participants collect and submit required data as a condition for enrolling in the program. The protocol also recommends the implementer specify the data reporting format, either by supplying a structured form (such as a spreadsheet) or by specifying the data fields and types used when submitting material to the program.

The format of the data must be electronic, searchable, and sortable. It must also support combining multiple files into single tables for analysis by the implementer. Microsoft Excel and comma-separated text files are acceptable formats; faxes, PDFs, and JPEGs are not.

The data reporting format should be structured to allow verification of the project installation. Each record or line in the report: (1) is a collection of identical fixture types, (2) is installed in an easily located room, floor, or space, and (3) belongs to one usage group. Table 1 lists the fields required in the data reporting format. All data are supplied by the participant or implementer.

Table 1. Required Lighting Data Form Fields

Field	Notes
Location	Floor number, room number, description
Usage group	
Location heating	Yes/no
Location heating type	Boiler steam/hydronic, rooftop gas-fired, etc.
Location heating fuel	Electric, natural gas, fuel oil, etc.
Location cooling	Yes/no
Location cooling type	Water cooled chiller, air cooled chiller, packaged DX, etc.
Location cooling fuel	Electric, natural gas, etc.
Baseline fixture type	From look-up table supplied by implementer, manufacturer cut sheet
Baseline fixture count	
Baseline fixture watt	From look-up table supplied by implementer, manufacturer cut sheet
Baseline HOU	From look-up table supplied by implementer, estimated by customer, bulk meter services or meter data
Efficient fixture type	From look-up table supplied by implementer, manufacturer cut sheet
Efficient fixture count	
Efficient fixture watt	From look-up table supplied by implementer, manufacturer cut sheet
Efficient lighting HOU	Same as baseline if no controls installed
IF_{cool}, or COP_{cool} and HOU_{cool}	Interactive factor for cooling from look-up table, or site-specific COP_{cool} and HOU_{cool} (optional)
IF_{heat}, or COP_{heat} and HOU_{heat}	Interactive factor for heating from look-up table, or site-specific COP_{heat} and HOU_{heat} (optional)
kWh_{save}	Calculated using savings algorithms
$kW\text{-}Peak_{save}$	Calculated using savings algorithms

The Appendix to this protocol contains an example of a lighting inventory form with the fields listed in Table 1.

Information at the usage-group level will typically not be available for midstream or upstream programs. Location-specific information such as cooling and heating type may also be unavailable. In such cases, program-level or building type-level defaults will be used by the implementer, and the evaluation may work to estimate and update these assumptions. See Section 8 for recommendations for the evaluation of midstream programs.

5 Role of the Evaluator

This section describes the evaluator's role in determining gross energy savings due to participation in a lighting energy efficiency program. Gross savings result directly from program-related actions taken by participants in an energy efficiency program. A simple way of thinking about gross savings is that they can be observed at the customer utility meter, at least in theory. In practice, it is difficult to isolate program-induced changes from other simultaneous changes, or to attribute them solely to the program itself.

Gross savings, the focus of this protocol, are adjusted through a separate set of actions to report net savings. Net savings are only those savings that can be attributed to the program. The concept of net savings recognizes that some participants would have acted on their own to adopt energy efficiency strategies, might have installed additional equipment as a result of increased awareness of the value of energy efficiency through their participation, or improved their energy efficiency operations due to market changes induced by a program's operations. For more on net savings, see UMP *Chapter 21: Estimating Net Savings – Common Practices*.

Steps taken by the evaluator under this protocol include:

1. Reviewing a statistically significant random sample of completed projects, including conducting on-site M&V activities
2. Calculating a gross realization rate (the ratio of evaluator-to-implementer anticipated gross savings)
3. Using the realization rate to adjust the implementer-estimated gross savings.

5.1 Evaluator Data Requirements

This protocol recommends the impact evaluator collect the same data as the implementer. As described in Section 6, the evaluator must have access to the implementation lighting inventory forms and participant application material for each project in the sample. For some program types, specifically midstream and upstream, the evaluator will collect more data than the implementer. This is the case when the evaluator conducts onsite verification of baseline conditions for a midstream program.

5.2 Evaluator Data Collection Method

Under the protocol, the implementer provides the evaluator with a copy of the program and project data tracking record for the evaluation review period. That record contains the fields specified in Table 1. The implementer also provides all records for projects in the evaluation review sample, including application materials and site contact information.

The protocol recommends the evaluator collect additional M&V data during site visits conducted for the sample of evaluation review projects. Table 2 lists data required for each project in the evaluation sample.

Table 2. Lighting Data Required by Evaluator

Field	Note
Location	From implementer
Usage group	From implementer
Location heating	From implementer, verified by evaluator
Location heating type	From implementer, verified by evaluator
Location heating fuel	From implementer, verified by evaluator
Location cooling	From implementer, verified by evaluator
Location cooling type	From implementer, verified by evaluator
Location cooling fuel	From implementer, verified by evaluator.
Baseline fixture type	From implementer, verified by evaluator
Baseline fixture count	From implementer, verified by evaluator
Baseline fixture watt	From implementer, verified by evaluator
Baseline HOU	From implementer, verified by evaluator
Efficient fixture type	From implementer, verified by evaluator
Efficient fixture count	From implementer, verified by evaluator
Efficient fixture watt	From implementer, verified by evaluator
Efficient lighting HOU	Measured by evaluator
IF_{cool}, or COP_{cool} and HOU_{cool}	Interactive factor for cooling from look-up table, or site-specific COP_{cool} and HOU_{cool} (optional)
IF_{heat}, or COP_{heat} and HOU_{heat}	Interactive factor for heating from look-up table, or site-specific COP_{heat} and HOU_{heat} (optional)
ISR	Measured by evaluator
kWh_{save}	Calculated using savings algorithms
$kW\text{-}Peak_{save}$	Calculated using savings algorithms

6 Measurement and Verification Plan

The M&V plan describes how evaluators determine actual energy savings in a facility where a lighting efficiency project has been installed. Evaluators use M&V to establish energy savings for a random sample of projects. The M&V results are applied to the population of all completed projects to determine program gross savings. The sampling and application processes are described in UMP *Chapter 11: Sample Design Cross-Cutting Protocol*. The sample size should be determined following the recommendations in UMP Chapter 11.

All M&V activities in the protocol are conducted on a representative sample of completed projects, drawn from a closed reporting period (for example, a program year).

6.1 IPMVP Option

The protocol recommends evaluators conduct M&V according to the International Performance Measurement and Verification Protocol (IPMVP) Option A—Retrofit Isolation: Key Parameter Measurement approach.

The key measured parameters are the HOU terms in Equation 1. The fixture quantity parameter is verified through an inspection process. The fixture wattage parameter is verified through a combination of on-site inspections and look-up tables of fixture demand (Watts).

Option A is recommended because the demand (Watts) values are known and published for nearly all fixture types and configurations, and therefore need not be measured, whereas lighting operating hours vary widely from building to building.

6.2 Verification Process

Verification involves visual inspections and engineering calculations to establish an energy efficiency project's potential to achieve savings. The verification process determines the fixture wattage and fixture quantity parameters in Equation 1.

The process includes the following steps:

1. Select a representative sample of projects for review (see UMP *Chapter 11: Sample Design Cross-Cutting Protocol* for guidance on sampling).
2. Schedule a site visit with a facility representative for each project in the sample.
3. Conduct an on-site review for each project. Inspect a representative sample of the energy efficiency lighting fixtures reported by the implementer. The protocol recommends selecting the sample from the implementer's inventory records before going on site (see UMP *Chapter 11: Sample Design Cross-Cutting Protocol* for guidance on sampling.)
4. Confirm or correct the reported energy-efficient fixture type and wattage for each fixture in the sample.
5. Confirm or correct the reported quantity for all energy-efficient fixtures in the sample.
6. Confirm or correct the heating/cooling status and associated equipment for the spaces in the sample.

7. Interview facility representatives to check baseline fixture types and quantities reported for the sample. Confirmation or correction is based on the interviews. When available, interviews are supplemented by physical evidence, such as: fixture types in areas not changed by the project, replacement stock for lamps and ballasts, and/or stockpiles of removed fixtures/lamps stored on-site for recycle or disposal.
8. Update lighting inventory form for the sample, based on findings from the on-site review.

The implementer has the primary responsibility for maintaining accurate project inventories that support evaluation research, including locations of individual fixtures or lamps. An example of an inventory form that meets the requirements of the protocol is provided in the Appendix.

Evaluators may have difficulty locating fixtures or lamps that contribute to a program's reported savings when project records are incomplete, information about fixture or lamp location and type is imprecise, the facility representative guiding the evaluator during a site visit is unfamiliar with the project, or the facility has undergone a change in ownership or retrofit since the project was completed.

When faced with incomplete information, evaluators can use a mix of strategies to conduct verification site visits. If a line in the inventory cannot be located, the verification sample can be expanded to include an entire floor, wing, or other space, and all of the fixtures within the space counted and identified. This approach works when room numbers are not provided in the inventory, for example. Another strategy is to substitute a room or space for one that cannot be located. This can work when there are large numbers of identical spaces such as classrooms, offices, and restrooms. Another is to contact the lighting contractor who installed the efficient lighting products. Contractors have to create their own inventories to manage construction and order material for all lighting projects, and they may be willing to share their lists.

Evaluation field staff will need to exercise judgement when using these strategies as to whether or not to count as verified any fixtures and lamps that cannot be located. The evidence can be inconclusive but still support a reasonable inference.

At the completion of the verification process, the evaluator has confirmed or corrected the fixture wattage and fixture quantity parameters in Equation 1. The process for determining the HOU parameters is described in the following section.

6.3 Measurement Process

The measurement process involves using electronic metering equipment to collect the data for determining the HOU parameters in Equation 1. Most often, the equipment is installed temporarily during the measurement period. Energy management systems that monitor lighting circuits can also be used to measure HOU.

Metering equipment used to measure lighting operating hours either records a change of state (light on, light off) or continuously samples and records current in a lighting circuit or light output of a fixture. All data must be time-stamped for application in the protocol.

6.3.1 Use of Data Loggers

Lighting operating hours are typically determined through the use of temporary equipment such as data loggers.

Change-of-state lighting data loggers are small (matchbox size) integrated devices, which include a photocell, a microprocessor, and memory. The data logger is mounted temporarily inside a fixture (or in proximity to it) and is calibrated to the light output of the fixture. Each time the lamp(s) in the fixture are turned on or off, the event is recorded and time-stamped.

Data loggers that continuously sample and record lighting operating hour information usually require an external sensor such as a current transformer (CT) or photocell. Data loggers with CTs can monitor amperage to a lighting circuit. Spot measurements of the circuit's amperage with the lights on and off establish the threshold amperage for the on condition. Similarly, a data logger with an external photocell can record light levels in a space. Spot measurements of lumen levels with the fixtures on and off establish the light level threshold for the on condition.

Although measuring amperage with data loggers is common, the continuous monitoring of lumen levels to determine hours of operation is less common.

Data logger failure commonly occurs due to incorrect adjustments, locations, or software launch. Thus, this protocol recommends following manufacturer recommendations carefully and deploying extra loggers as a cushion against failure.

6.3.2 Metering

The measurement process involves metering lighting operating hours for the representative sample of fixtures selected for the verification process. Meters are deployed or trends set up in an existing energy management system during the verification site visit.

This process entails the following activities:

1. Meter operating hours for each circuit in the verification sample.
    A. Develop a metering plan that includes the location of a random selection of required metering points (the metering sample) by usage group. Guidance on sampling is provided in UMP *Chapter 11: Sample Design Cross-Cutting Protocol*. The plan should be developed before going on site. If the inventory is missing good location information, the plan can be adjusted while in the field to make sure the number of metering points by usage group is maintained and that the selection remains random to the extent possible.
    B. If using light loggers, deploy loggers in one or more fixtures controlled by the circuit. Only one logger is required per circuit; additional loggers may be deployed to offset logger failure or loss. A rule of thumb is to install the number of loggers specified in the metering plan for each usage group plus an additional 10%.
    C. If measuring amperage, install CT and data logger in a lighting panel for a sampled circuit. The sampling interval should be 15 minutes or less. Spot-

measure amperage with lights on and off for the circuit leg with the CT. Record the amperage threshold for the lights-on condition.

   D. If using an energy management system, program it to sample and record lighting on/off status for each circuit in the sample. The sampling interval should be 15 minutes or less. Check that the energy management system has sufficient capacity to archive recorded data, and that the metering task will not adversely slow system response times.

2. Check data logger operation. Before leaving the site, spot-check a few data loggers to confirm they are recording data as expected. Correct any deficiencies and if the deficiencies appear to be systemic, redeploy the loggers. If using energy management system trends, spot-check recorded data.

3. Leave the metering equipment in place for the duration of the monitoring period. The protocol recommends a monitoring period that captures the full range of facility operating schedules. The following are some rules of thumb for specifying the length of the monitoring period. More detailed guidance is provided Section 6.3.3.

   A. For facilities with constant schedules (such as office buildings, grocery stores, and retail shops), the protocol requires metering for a minimum of four weeks. The weeks should not be abnormal (e.g. during the end-of-year holidays).

   B. For facilities with variable or irregular schedules, additional metering time is required. The protocol recommends a monitoring period long enough to capture the average operation over the full range of variable schedules.

   C. Facilities with seasonal schedules, such as schools, should be monitored during active periods; additional monitoring can be done during the inactive periods, or if the expected additional savings are small, the hours can be estimated as a percent of active period hours.

4. Analyze metering data. Calculate the percent on-time for the metered lighting equipment for each usage group. Percent on-time is the number of hours the lighting equipment is on divided by the total number of hours in the metering period. Annual lighting hours are the percent on-time times 8,760 hours per year less any closed hours such as for holidays. Separate on-time factors can be developed for day-of-week, month-of-year, and seasonal timeframes if the metered data capture the full range of operations for the more granular reporting period.

   A. For facilities with constant or variable schedules, the HOU parameter is calculated as: 8,760 hours per year, less any hours when the facility is closed for holidays, times the percent-on time.

   B. For facilities with seasonal schedules, the HOU parameter is: the hours/year in the active or operational period, times the percent-on time.

   C. The data used in the analysis should represent a typical schedule cycle. For example, 28 full days for an office space occupied Monday through Friday and unoccupied on weekends. The hours/year in the active period may vary by usage group; in schools, for example, office spaces may be active 8,760 hours/year, while classrooms are only active 6,570 hours/year.

5. Evaluation timing usually requires the measurement of operating hours after the efficiency project has been completed. This process assumes that the operating hours are unchanged from the baseline period. Thus, HOU baseline and HOU energy-efficient in Equation 1 have the same value. (Note that will not be the case if the project includes lighting control measures.)
6. UMP *Chapter 3: Commercial and Industrial Lighting Controls Evaluation Protocol* addresses lighting control measures, but Equation 1 can accommodate changes in lighting operating hours, as would occur in combined lighting equipment and lighting controls projects, provided measured hours of use data are available for the baseline period. For example, these data may be available for a facility with an energy management system with archived trends or if a lighting contractor conducted a metering study before entering into a performance contract.

6.3.3 Duration of Metering Period

While a metering period of one year would provide the most accurate picture of a facility's lighting HOU, economic and customer considerations impose practical limits on the actual duration. Regulators, program administrators, and customers have limits to their tolerance for lengthy evaluation periods that delay studies and their results, and that require on-going facility coordination. Evaluators are thus faced with the questions, "What is the optimal length of a metering study to obtain acceptable estimates of annual lighting hours of use?" and "How accurate is this optimal estimate?" A recent study conducted for a Massachusetts large commercial/industrial program provides some answers (KEMA 2013).

This protocol recommends a one-month minimum metering period based, in part, on the results of this long-term Massachusetts study.

The study included 12 months of continuous monitoring of lighting systems at 34 large commercial and industrial sites. Evaluators estimated gross annual savings from each month of data collected, as well as for each two- and three-month block of data collected. The three-month results were later compared to the full 12-month study results to determine how well these shorter metering periods results following completion of the full year of monitoring. The key findings were that a three-month period of monitoring did a reasonable job of estimating full year savings as compared to the 12 months of monitoring. **Error! Reference source not found.** below presents the HOU and summer coincidence factors by building type from both monitoring periods. A value greater than 100% means that the full year was higher than the three-month estimate, or that the three-month data underestimated these parameters. Due to the seasonal usage, school/university-type buildings were more difficult to annualize and estimate summer coincidence factors from three months of data.

For most sites, the three-month period included winter/spring months. Fewer daylight hours in the winter as compared to summer in the northern hemisphere explain why the three-month results overestimated HOU and CF for this building type as compared to the full 12-month study.

Table 3. Comparison of 12-Month and 3-Month HOU and Summer CF by Building Type

Building Type	Count of Building Type	3-Month Data - Annualized HOU	Actual 12-Month Data HOU	12-Month/ 3-Month HOU	3-Month Data - Estimated Summer CF	Actual 12-Month Data Summer CF	12-Month/ 3-Month CF
Manufacturing (n=6)	6	5,898	5,730	97%	88%	88%	100%
Office (n=5)	5	4,079	3,759	92%	89%	81%	91%
Retail (n=5)	5	5,727	5,473	96%	91%	91%	100%
School/University (n=4)	4	3,114	2,839	91%	54%	39%	72%
Exercise Center (n=2)	2	6,541	6,604	101%	89%	91%	102%
Library (n=2)	2	2,129	1,990	93%	58%	58%	101%
Other (n=10)	10	6,054	5,965	99%	81%	79%	98%
All Lighting Systems (n=34)	34	5,140	4,963	97%	81%	77%	96%

The study also looked at the differences in annual energy savings when using blocks of one-, two-, and three-month metering periods compared to the 12-month study results, as summarized in Table 4. The percentages represent how close the annual energy savings would have been compared to the full 12-month results had data from each specified that period been used to estimate annual energy savings. Each of the monitoring periods were able to produce annual energy savings estimates to within 10% of the full 12-month result, which is the basis for the protocol's recommendation for a one-month metering minimum. As more data were included, the annual savings estimates improved. By including three months of data, evaluators could estimate annual energy savings to within 5% of the full 12-month result regardless of the specific three-month period.

Table 4. Comparison of 1-, 2-, and 3-Month Metering to Actual 12-Month Energy Savings

Monitored One Month	Percent of Actual Annual Energy Savings	Monitored Two Month Period	Percent of Actual Annual Energy Savings	Monitored Three Month Period	Percent of Actual Annual Energy Savings
January	99%	Jan-Feb	95%	Jan-Mar	97%
February	90%	Feb-Mar	95%	Feb-Apr	96%
March	101%	Mar-Apr	98%	Mar-May	101%
April	96%	Apr-May	102%	Apr-Jun	102%
May	108%	May-Jun	106%	May-Jul	104%
June	103%	Jun-Jul	103%	Jun-Aug	103%
July	102%	Jul-Aug	103%	Jul-Sept	102%
August	104%	Aug-Sept	103%	Aug-Oct	103%
September	101%	Sept-Oct	103%	Sept-Nov	100%
October	105%	Oct-Nov	99%	Nov-Jan	97%
November	94%	Nov-Dec	96%	Oct-Dec	99%
December	98%	Dec-Jan	98%	Dec-Feb	96%

If either winter or summer peak demand savings are of concern, the protocol recommends including at least one winter or one summer month in the monitoring period. If both winter and summer peak demand savings are equally important, the protocol recommends monitoring during both seasons. Note that monitoring for both seasons extends the study timeline to at least nine months and increases the overall cost.

6.4 Report M&V and Program Gross Savings

Information collected during the M&V processes is used to calculate M&V project savings, as follows:

1. Using the results from the last step in verification process, update the inventory HOU parameters and calculate M&V savings for the sample of projects.
2. Calculate the program gross realization rate, the verified project savings divided by the reported project savings for the sampled projects.

Equation 10. Program Gross Realization Rate

$$Gross\ Realization\ Rate_{kWh,kW} = \frac{\sum kWh, kW_{M\&V}}{\sum kWh, kW_{Reported}}$$

3. Calculate the evaluated program savings, the product of the program realization rate and the program reported savings.

Equation 11. Evaluated Program Savings

$$Evaluated\ Savings_{kWh,kW} = Gross\ Realization\ Rate_{kWh,kW} \cdot kWh, kW_{Reported}$$

The uncertainty and, therefore, the reliability of the program realization rate depend on the sample size and variance in the findings (described in *Chapter 11: Sample Design Cross-Cutting Protocol*). These are usually a function of the confidence and precision targets stipulated by regulators or administrators, and evaluation budgets. The sample sizes for homogeneous lighting efficiency programs can range from as few as 12 for an 80/20 confidence/precision target to as many as 68 (or more) for a 90/10 target, assuming an average coefficient of variation of 0.5. Higher coefficient of variations will result in larger samples.

The confidence level and its associated precision of the evaluated savings in

Equation 11 should be included when reporting results; for example, 732 MWh/year ±7% (relative), or 732 MWh/year ±51.2 MWh/year (absolute) at 90% confidence. UMP *Chapter 11: Sample Design Cross-Cutting Protocol* describes the calculation of precision for reported savings. A worked example showing the precision calculations for reported savings from a lighting project is also available as part of the IPMVP.[4]

Precision can only be calculated for the metering period. Care should be taken to ensure that the metering period is representative of the entire year as described in Section 6.3.3.

6.5 Data Requirements and Sources

This section contains information on the fixture wattage, annual HOU, interactive cooling, and interactive heating factor parameters found in the algorithm equations. Data requirements are described in Section 4 *Role of the Lighting Program Implementer* and Section 5 *Role of the Evaluator*, with additional detail in Section 6 *Measurement and Verification Plan*.

6.5.1 Fixture Wattage

The protocol recommends use of fixture wattage tables, developed and maintained by existing energy efficiency programs and associated regulatory agencies. The tables list all common fixture types. Most tables are updated as new fixtures and lighting technologies become available.

The wattage values are measured according to ANSI standards[5] by research facilities working on behalf of manufacturers and academic laboratories (CEC 1993).

In the wattage table, each fixture and screw-in bulb is fully described and assigned a unique identifier. The implementer enters a fixture code into a lighting inventory form, which, if programmed, can search by a look-up function to show the associated demand. The evaluator

[4] IPMVP, Uncertainty Assessment. Anticipated to be available to the public winter of 2018. http://evo-world.org/en/

[5] The ANSI 82.2-2002 test protocol specifies ambient conditions for ballast/lamp combinations in luminaires. The test is conducted on an open, suspended fixture. Actual fixture wattage will vary, depending on the installation (suspended, recessed) and housing type. Differences are small—less than 5%.

then verifies or corrects the fixture type for the evaluation sample, updating the lighting wattage values if needed.

Fixture wattage tables do not include records for many LED fixtures and lamps in part because the tables lag behind this developing technology, but also because LEDs do not lend themselves to the same clear-cut classifications used for older technologies such as fluorescent or metal halide. LED fixture codes are needed to classify them by application and cost so they support market trending and cost-effectiveness analysis.

A solution is to allow users to create LED fixture codes that capture type and wattage using a set nomenclature. Following is an example of one scheme:

LEDnXXXXww

Where:

n = number of lamps

XXXX = fixture category from Table 5

ww = fixture wattage from manufacturer cut sheets

Table 5. LED Fixture/Lamp Categories

48" Linear Fluorescent Tube Replacement	LT
24" Linear Fluorescent Tube Replacement	LT
High-Bay Luminaires	HBR
Outdoor Pole/Arm-Mounted Luminaires	OP/A
Outdoor Wall-Mounted Luminaires	OW
Refrigerated Display Case Luminaires	RDL
Street Lamp Luminaires	ST
Custom	C

Thus the fixture code LED1OP/A50 for a 10-lamp, 50-watt outdoor pole/arm-mounted luminaire.

The protocol recommends adopting a fixture wattage table, used by an established and recognized lighting efficiency program. As of August 2017, the following sources provide examples (many others are available in most U.S. regions):

- Massachusetts Program Administrators. (October 2011). *Massachusetts Technical Reference Manual for Estimating Savings from Energy Efficiency Measures* (October 2015). http://ma-eeac.org/wordpress/wp-content/uploads/2016-2018-Plan-1.pdf.
- TecMarket Works. (October 2010). *New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs—Residential, Multi-Family and Commercial/Industrial Measures*. (Version 5) (July 2017). Prepared for the New York

Public Service Commission. http://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/72c23decff52920a85257f1100671bdd/$FILE/TRM%20Version%205%20-%20January%202018.pdf.

- Database for Energy Efficiency Resources (DEER). Available from the California Public Utilities Commission at: www.deeresources.com. An exhaustive list of all parameters driving energy use and savings for a lengthy list of measures. References California codes and weather zones.

Wattage tables are used by both the implementer and the evaluator. An excerpt from the *New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs* is included in the Appendix to this protocol as an example.

6.5.2 Hours of Use

The protocol requires the evaluator to measure operating hours for a sample of buildings and fixtures, as described in Section 6.3 *Measurement Process*.

This section describes data sources and methods used by the program implementer for estimating HOU values for individual projects. Accurate estimates of the HOU parameter are needed for the implementer to reliably estimate and report project and program savings. Accurate reporting by the implementer also results in more accurate evaluated savings for a given sample size.

The protocol requires program participants to provide estimates of HOU values by usage group in their lighting inventory forms. The estimate should not be based on the building schedule alone, although this may inform the estimate. Instead, the protocol recommends participants develop the HOU values using one of the following sources, with guidance from the program implementer:

- ***Lighting schedules in buildings with energy management systems or time clocks*** controlling lighting equipment. The project participant should interview the building manager to verify that the schedules are not overridden. Control schedules (or trend data) are reliable estimates of true lighting operating hours, but they are normally available only for larger, newer facilities.

- ***Interviews with building managers.*** Building managers are usually familiar with lighting schedules, and can describe when lights are turned on and off for typical weekdays and weekends. They may not know about abnormalities such as newly vacant spaces, how cleaning crews operate lights, or whether lights are actually turned off after hours. The protocol recommends interviewing two or more people familiar with a facility's operation to verify scheduling assumptions.

- ***Tables of HOU values by building type*** provided by the program implementer. HOU values have been developed from impact evaluation and M&V studies for many commercial and nonresidential buildings. Like wattage tables, HOU tables are maintained by energy efficiency programs and associated regulatory agencies; sources can be found using the same references provided for wattage tables. An excerpt from the *New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs* is included in the *Appendix* to this protocol as an example of a table of HOU values.

Actual operating schedules vary widely for any given building type, and tabulated average values contain larger variations than values for fixture wattages. Also, tabulated HOU values are given for entire buildings, not by usage groups within buildings. The protocol requires HOU estimates be entered into the inventory by usage group, which will vary from the building average. For these reasons, the protocol recommends use of building-specific lighting operating hours when these are available, supplemented if necessary by tables of HOU values.

6.5.3 Interactive Cooling and Heating

Energy-efficient lighting equipment contributes less waste heat to building conditioned spaces, compared to baseline equipment. This results in a reduced cooling and increased heating loads.

This protocol provides two options for calculating interactive cooling and heating effects: an engineering approach and a stipulated factor approach. A third approach, simulation modeling, is also an option; however, it tends to be labor-intensive and is usually reserved for large-scale studies used to quantify stipulated factors. It is unusual to model interactive savings on a project-by-project basis, and it is not required by the protocol.

The engineering approach requires site-specific estimates of the COP for the cooling and heating equipment, and the lighting HOU coincident with the cooling or heating equipment operation. These values can be developed from information gathered during site visits conducted as part of the verification process.

6.5.3.1 Interactive Cooling and Heating – Stipulated Approach

The stipulated factor approach uses interactive factors—terms IF_{cool} and IF_{heat} in Section 3.1 *Algorithms*—to account for the additional changes in cooling or heating energy use. Values are dependent on the type of facility, regional climatic conditions, and cooling and heating equipment. Guidance is provided below for several common situations.

Interactive cooling effects are generally small for spaces conditioned for human comfort (2% to 6% for cooling in offices in New York City, for example) (TecMarket Works 2010). They are also highly dependent on HVAC system types and efficiencies. For example, in a large office building in New York City, the IF_{cool} varies with the equipment: (1) with gas heat and no economizer, the IF_{cool} is 3.3%, (2) with an economizer, the IF_{cool} is 1.9%, and (3) with economizer and a variable air volume system, the IF_{cool} is 6.5%. In regions with hot climates where cooling loads are higher than in New York City, IF_{cool} values will be larger than these examples. In cooler climates, the values will be lower.

Interactive heating effects are also small for conditioned spaces and will vary with HVAC system types and efficiencies. For example, in a large office building in New York City, the IF_{heat} ranges from -2.2% to -1.3% (TecMarket Works 2010). The negative value indicates that decrease in waste thermal energy from the efficient lighting equipment must be replaced by the heating system.

Electric efficiency programs often ignore interactive heating effects when territory's heating systems are primarily nonelectric; e.g., natural gas or oil. For comprehensive programs with an all-fuels reporting responsibility, or where electric heating is significant, the increased heating energy can be included.

Interactive factors are usually too small to be measured accurately; instead, they are developed using computer simulations and the interactive impacts are stipulated. Interactive effects are available from the same sources as fixture wattages and HOU.

Interactive effects can be significant in cold-temperature conditioned spaces, such as freezers or refrigerated warehouses. For example, in Pennsylvania, the default interactive cooling factors are defined by space temperature ranges as follows (Pennsylvania Public Utility Commission 2016):

- Freezer spaces (-20 °F–27 °F) = 50%
- Medium-temperature refrigerated spaces (28 °F–40 °F) = 29%
- High-temperature refrigerated spaces (47 °F–60 °F) = 18%
- Uncooled space (e.g. warehouse with no mechanical cooling) = 0%.

Not all programs estimate, report, and evaluate interactive effects, and the decision is often a policy choice. Further, because programs are often energy specific (electricity or gas), the effect on other fuels is sometimes ignored. For example, electric energy efficiency programs might report interactive electric cooling savings, but omit interactive increases in gas heating energy.

A sample of IFs can be found in the documents listed in *Resources*.

6.5.3.2 Interactive Cooling and Heating – Engineering Approach

A complete description of the engineering approach to estimating interactive cooling and heating effects is provided in Section 3.1.1 *Energy Savings.*

6.5.4 Coincidence Factors (CF)

CFs adjust the change in connected electric load from lighting efficiency projects for electric peak demand savings. Electric demand savings that occur during utility system peak periods help lower utility capacity requirements, reduce the load on peak generation equipment that is usually the costliest to operate, and improve system reliability. The value of peak demand generation is reflected in rate structures that charge customers for their demand during peak time-of-use periods.

CFs can range from a high of 1.0 down to 0.0, where 1.0 indicates that 100% of a lighting project's change in connected load occurs during the utility peak period. An example is the CF of 1.0 for commercial lighting efficiency projects in New York State (TecMarket Works 2010). Dawn-to-dusk exterior lighting has a CF of 0.0 when system peaks occur during daylight hours, which is normal for most utilities. Some programs or utilities may have very specific targets for the timing of demand reductions. For example, the Con Edison Brooklyn Queens Demand Management Program targets savings from 9 p.m. to 10 p.m. on weekdays. Use of typical commercial CF for such a program is not advised.

CFs can be developed from lighting HOU meter data. The CF is the peak period energized lighting kW as measured by the meter data, divided by the total connected kW for the energy efficiency lighting project.

This protocol recommends using tables of CFs (including any interactive effects from reduced cooling loads) to report system peak coincident electric demand savings. If regulators or program administrators require greater reliability for evaluated demand reductions (as would occur for a program designed to increase capacity reserves), CFs should be developed from metered data. Like IFs, unique CFs can also be adapted from programs with similar customer and utility profiles.

A sample of CFs can be found in the documents listed in *Resources*. CFs are also discussed in UMP *Chapter 10: Peak Demand and Time-Differentiated Energy Savings Cross-Cutting Protocol*.

7 Gross Impact Evaluation

Gross impact evaluations entail a detailed review of a random sample of completed projects, concluding with an independent assessment of their gross savings. The ratio of program-claimed savings and gross evaluated savings for the projects (the gross realization rate) is used to adjust claimed savings for all completed projects (the program).

Gross impact evaluations are coordinated in conjunction with program milestones, usually at the end of a program year or cycle. The evaluation's subject is the population of all projects completed up to the milestone.

It is preferable to begin evaluation activity before the program cycle ends, because difficulties and inaccuracies often occur when collecting data retroactively, particularly in attempts to backfill missing data, determine baseline data, or deal with poor customer recall of project details. This may require drawing a preliminary sample before the milestone date and then adjusting (adding to) the sample after the milestone date.

The evaluator uses the same algorithms and data as the program implementer (subject to review and site inspections), except that HOU values are based on measurements of actual lighting operating hours for all projects in the evaluation sample, and lighting inventories (including baseline and energy efficiency fixture types and counts) are corrected as needed based on on-site reviews of the sample projects.

The ratio of evaluator savings to program reported savings for the projects in the M&V sample is the program realization rate. Total reported program savings for the reporting period are then multiplied by the program realization rate to determine program evaluated savings for the period.

Realization rates can also be developed for facility and customer types if the implementer is interested in the savings performance of these sub-populations.

7.1 Sample Design

The protocol requires sampling to select:

- Projects from a program database for an impact study
- Inventory lines for deploying light loggers.

Regulators normally prescribe the confidence and precision levels for the sample, or the implementer may impose them. UMP *Chapter 11: Sample Design Cross-Cutting Protocol* describes general sampling procedures and should be consulted when developing evaluation plans for lighting efficiency programs. The following details pertain specifically to lighting.

The protocol recommends stratified sampling when selecting projects for an impact study because it usually results in smaller sample sizes as compared to simple random sampling. The idea behind stratified sampling is to select subpopulations of relatively homogeneous projects such that the variance within each stratum is smaller than for the population as a whole, as explained in UMP *Chapter 11: Sample Design Cross-Cutting Protocol*.

A simplified stratified strategy is to rank all projects in the population to be studied by their reported savings (ranked from largest to smallest) and to define three strata. The top stratum contains large projects that cumulatively account for 50% of reported savings, and the remaining projects are grouped into medium strata contributing 30% and small strata contributing 20%.

A more rigorous method is to use a stratified strategy with a customized stratum threshold where techniques are employed to define strata that minimize the expected variance in their realization rates, and thereby minimize the sample size. The stratification thresholds are designed to minimize the variance of a stratified ratio estimator. Stratified ratio estimation is fully explained in UMP *Chapter 11: Sample Design Cross-Cutting Protocol*, which should be referenced when developing sampling plans. Projects may also be stratified by technology types, or by other characteristics, if known, such as business type or primary space type of the installations.

Light-logger studies also use stratified sampling within projects selected for M&V by selecting samples of fixtures for metering, with strata defined by usage groups. The desired confidence and precision interval (typically prescribed with an assumed coefficient of variation of 0.5) determines the sample size. The Federal Energy Management Program M&V Guidelines (Federal Energy Management Program 2008) describe a detailed routine for selecting logging lines.

Oversampling of projects by 30% and of loggers within projects by 10% is recommended to replace participants that cannot be scheduled for a site visit, and to provide a cushion against lost or failed loggers in HOU studies.

8 Other Evaluation Issues

8.1 Upstream/Midstream Delivery

As upstream programs do not interact with individual customers, they lack the lighting inventory forms (with associated data) used to estimate savings. Implementers and evaluators can use sales data, surveys, saturation studies, and other indirect methods to estimate baseline fixture wattages and facility HOUs. Implementers and evaluators can also draw on incentive and rebate program data by analyzing baseline fixtures and operating hours associated with fixtures promoted in the upstream buy-down program, thereby developing savings factors for upstream buy-down equipment.

Midstream programs are a subset of the upstream family where incentives for qualified lighting products are paid to distributors selling to contractors and facility managers. Implementers can leverage the distributor-purchaser relationship to collect key information needed for evaluation. This information includes the purchased equipment and the site where it will be installed. Many of the details such as baseline equipment, scheduling, and lighting HOU for the installation site facility must be collected after the sale, by the evaluator, on a random sampling basis. The implementer must make assumptions for these and deem them to report savings. Table 6 lists data required for each project in an evaluation sample, and shows the source of each element.

Table 6. Lighting Data Required by Evaluator for Midstream Programs

Field	Data Source	
	Implementer	Evaluator
Facility	Distributor invoice	From implementer
Facility type	Distributor or utility account	Evaluator data gathering
Usage group	Not reported	Evaluator data gathering
Facility heating (yes/no)	Deemed	Evaluator data gathering
Facility heating type	Deemed	Evaluator data gathering
Facility heating fuel	Deemed	Evaluator data gathering
Facility cooling	Deemed	Evaluator data gathering
Facility cooling type	Deemed	Evaluator data gathering
Facility cooling fuel	Deemed	Evaluator data gathering
Baseline fixture type	Deemed based on efficient fixture type	Evaluator data gathering
Baseline fixture count	Deemed based on efficient fixture count	Evaluator data gathering
Baseline fixture watt	Deemed based on efficient fixture type	Evaluator data gathering
Baseline HOU	Deemed based on facility type	Evaluator data gathering
Usage group	Not reported	Evaluator data gathering
Efficient fixture type	Distributor invoice	From implementer
Efficient fixture count	Distributor invoice	From implementer
Efficient fixture watt	Distributor invoice, qualified products list	From implementer
Efficient lighting HOU	Deemed, look-up table by facility type	Measured by evaluator
IF_{cool}, or COP_{cool} and HOU_{cool}	Deemed, look-up table by facility type	Interactive factor for cooling, from look-up table or evaluator data gathering, optional
$IF_{heat,}$ or COP_{heat} and HOU_{heat}	Deemed, look-up table by facility type	Interactive factor for heating, from look-up table or evaluator data gathering, optional

Field	Data Source	
	Implementer	Evaluator
ISR	Deemed based on previous studies	Evaluator data gathering
kWh_{save}	Calculated using savings algorithms	Calculated using savings algorithms
$kW\text{-}Peak_{save}$	Calculated using savings algorithms	Calculated using savings algorithms

8.1.1 Role of the Implementer in Lighting Midstream Programs

Successful application of this protocol to midstream lighting programs requires collecting standard data in a prescribed format as part of the implementation process. The protocol further requires tracking project and program savings estimated on the basis of those standard data.

Distributors are required to submit sales invoices to the implementer. Invoices capture the efficient lighting product type, specifications, and quantity for each purchase.

Because the implementer does not have contact with end-users who purchase efficient lighting products through midstream programs, savings estimates must make assumptions for five baseline variables used in Section 3.1 equations. Following are standard approaches to determining the baseline assumptions for the five variables used to report program savings. The assumptions and savings are subject to revision by an evaluation review.

1. **Baseline fixture/lamp wattage.** Most programs will use a replace-on-burnout baseline where existing equipment would fail and likely be replaced by a minimally code- or standard-compliant product, or the most commonly installed product if not regulated. Thus, the implementer will match each efficient product with a baseline specification using codes and standards, or market practice. For example, a four-foot LED tube is assumed to replace a T8 lamp. Another example, high bay LED fixtures can be mapped to high-intensity discharge (HID) fixtures using lumen bins; a 15,500 to 20,100 high bay LED replaces a 462-watt (400-watt lamp) metal halide fixture. An example of table-mapping high bay LED fixtures to baseline HID fixtures is provided in the Appendix. In midstream programs, the evaluator can determine, based on site visits, if an early replacement baseline should be used rather than replace-on-burnout.

2. **Baseline fixture/lamp quantity.** Assume a one-for-one replacement. Baseline quantity is equal to the efficient product quantity.

3. **Annual HOU.** Identify the building type where the efficient product is installed and use look-up table to select HOU values by building type. The building type can be identified by using business name, address, and utility account number for each sale, or by requiring distributors to collect it at the point of purchase.

4. **Interactive cooling factor, or cooling equipment COP.** Use look-up table of deemed interactive factor or equipment COP by building type.

5. **Interactive heating factor, or heating equipment COP.** Use look-up table of deemed interactive factor or equipment COP by building type.

The implementer uses the invoice and assumed baseline data to report savings using the equations in Section 3.1.

The implementer is responsible for ensuring all necessary data are collected to track program activity and to calculate savings at the project level. The implementer is responsible for maintaining a program activity record, including anticipated savings by project.

8.1.2 Role of the Evaluator in Lighting Midstream Programs

As described in Section 5, the evaluator's role in midstream programs is to determine energy savings resulting from the operation of lighting efficiency programs. The unique feature for midstream programs is the need to collect more of the baseline and facility data for each project in the evaluation sample as indicated in Table 6. The steps in this procedure include:

1. Reviewing a sample of completed projects, including conducting on-site M&V activities
2. Calculating a gross savings realization rate (the ratio of evaluator-to-implementer anticipated savings)
3. Using the realization rate to adjust the implementer-estimated savings.

8.1.2.1 Evaluator Data Requirements

The protocol recommends that the program evaluator develop the same data as the implementer. However, for midstream programs, the sources for most of the data points will be different for each party; the implementer is forced to make assumptions for the baseline and facility while the evaluator contacts each facility in the evaluation sample to verify the actual conditions. The evaluator must have access to the distributor sales data for the sample of incentivized lighting products, including information about the facility where they are installed.

8.1.2.2 Evaluator Data Collection Method

Under the protocol, the implementer provides the evaluator with a copy of the program and project data tracking record for the evaluation review period. That record contains the fields specified in Table 6. The implementer also provides all records for projects in the evaluation review sample, including application materials and site contact information.

This protocol recommends the evaluator collect M&V data during site visits conducted for the sample of evaluation review projects. The data include information about the baseline equipment using the same techniques as for rebate and incentive programs. The data are used to update assumptions and values made by the implementer. Table 6 lists data required for each project in the evaluation sample.

8.2 New Construction

Installed power (kW) savings for new construction projects are calculated by subtracting as-built building lighting power from the lighting power of a code-compliant alternative, or common practice. The code-compliant alternative or common practice is the baseline. For jurisdictions where common practice is more efficient than code, a common practice baseline should be used. This is occurring in regions where LED lighting is specified in new construction (as opposed to T8/CFL/Metal Halide technologies). Code defines compliance in terms of lighting power density (LPD, lighting watts/ft^2). Lighting power equals LPD times building area.

New construction codes require controls with automatic lighting shutoff, with some exceptions for safety. The controls reduce the lighting HOU compared to existing facilities. Implementers can use look-up tables of new construction lighting hours of use that account for controls. These tables are available from some of the references in Section 11. Evaluators measure HOU using meters or bulk meter services.

8.3 First Year Versus Lifetime Savings

This protocol provides planners and implementers with a framework for reliable accounting of energy and demand savings resulting from lighting efficiency programs during the first year of measure installation.

Savings over the life of a measure may be less than the product of first-year savings and measure life. The discount results from replacement, degradation, or failure of the efficient equipment. Lifetime savings are covered further in UMP *Chapter 13: Assessing Persistence and Other Evaluation Issues Cross-Cutting Protocol*. However, because lifetime savings for lighting projects are strongly driven by federal standards and changes in the market, they are discussed here.

Most T12 lamps do not meet federal efficacy (lumens/watt) standards that went into effect in July 2012, accelerating a long-term trend toward T8 and T5 lamps and electronic ballasts, or LED tubes or panels. The effect is that first-year savings for T12 to T8, T5, and LED replacements can be assumed only for the remaining useful life of T12 equipment, at which point customers have no choice but to install equipment meeting the new standard.

For retrofit lighting programs, at the time when old equipment would be replaced, there is effectively a step up in the baseline and a step down in the annual savings for the replacement equipment. This leads to a dual baseline:

- An initial baseline with full first-year savings for the remaining useful life of the replaced equipment
- An efficient baseline with reduced savings for the remaining effective useful life of the efficient equipment.

The protocol methodologies, which specify tracking data for each installation, support the calculation of lifetime savings (including the use of a dual baseline).

The remaining useful life can be estimated from research studies. It can also be assumed to be a third of the effective useful life of the baseline equipment.

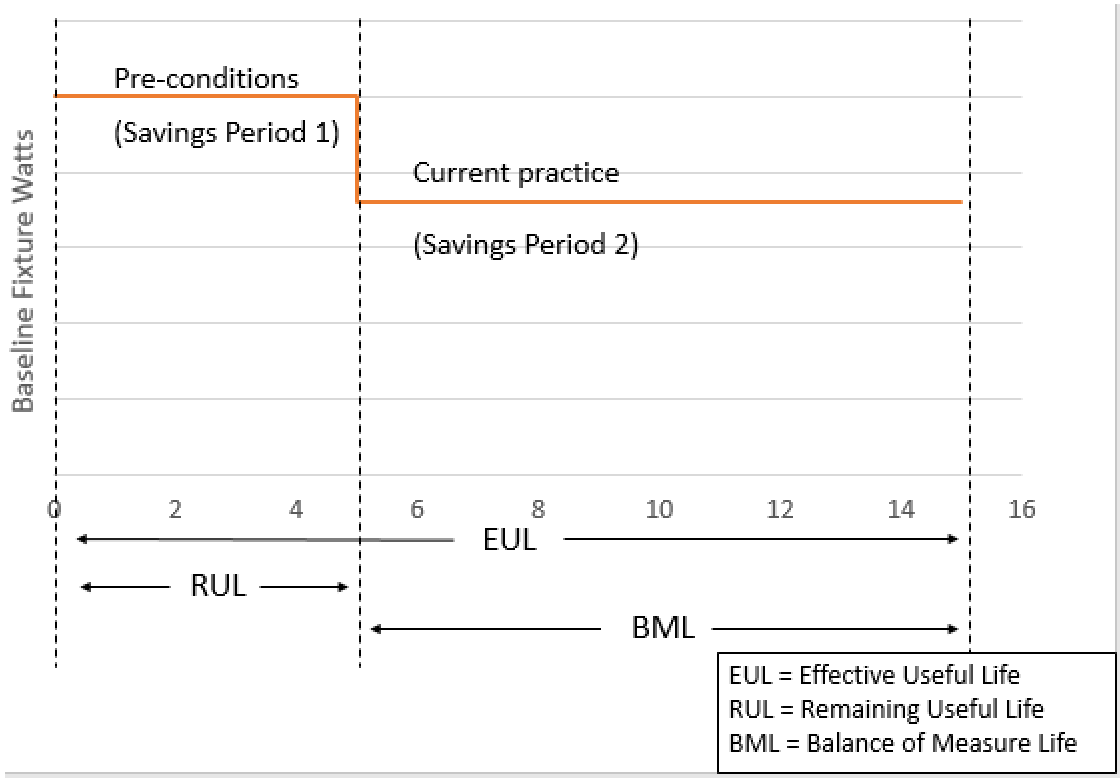


Figure 1. Dual baseline[6]

Figure courtesy of Regional Technical Forum

8.4 Program Evaluation Elements

Setting the foundation for a successful evaluation of a commercial, industrial, non-residential lighting program begins early in the program design phase. Implementers support future evaluations by ensuring data required to conduct an impact study are collected, stored, and checked for quality. These data include measured and estimated values available from past studies or equipment tests. Implementers must set data requirements before a program's launch to ensure the information required to conduct the research will be available.

[6] "Current practice" in the "Savings Period 2" time frame could be codes and standards, or current market practice for products not covered by codes or standards.

9 Looking Forward

Market baseline studies can support gross impact evaluation research of upstream programs by identifying associations between incentivized products and categories of baseline equipment, including their demand and energy patterns. Longitudinal market effects studies can supplement traditional site visit data gathering by characterizing changes over time in lighting equipment installations.

There is a need to develop hybrid approaches for lighting programs that include both market baseline and market effects studies in addition to the sampling and site visit model described in this protocol. As the delivery of lighting energy efficiency changes to include upstream and midstream models along with traditional downstream (rebate) models, as appears to be occurring now, there will be greater need for these market data to 1) establish baselines and 2) quantify gross (and net) savings impacts.

10 References

Chantrasrisalai, C.; Fisher, D.E. (2007). Lighting heat gain parameters: Experimental results. HVAC&R Research 13(2):305-324.

California Public Utilities Commission (CPUC). Database for Energy Efficient Resources (DEER). Available online: www.deeresources.com

Federal Energy Management Program (FEMP). (2008). *M&V Guidelines: Measurement and Verification for Federal Energy Projects Version 4.0*. https://energy.gov/sites/prod/files/2016/01/f28/mv_guide_4_0.pdf.

KEMA, Inc. (June 21, 2013). *Impact Evaluation of 2010 Prescriptive Lighting Installations*. Prepared for the Massachusetts Program Administrators and Massachusetts Energy Efficiency Advisory Council. http://ma-eeac.org/wordpress/wp-content/uploads/Impact-Evaluation-of-2010-Prescriptive-Lighting-Installations-Final-Report-6-21-13.pdf

Pennsylvania Public Utility Commission. (2016). *Technical Reference Manual*, Appendix C. http://www.puc.state.pa.us/filing_resources/issues_laws_regulations/act_129_information/technical_reference_manual.aspx.

TecMarket Works. (July 2017). *New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs—Residential, Multi-Family and Commercial/Industrial Measures*. (Version 5). Prepared for the New York Public Service Commission. http://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/72c23decff52920a85257f1100671bdd/$FILE/TRM%20Version%205%20-%20January%202018.pdf

11 Resources

This protocol depends heavily on reliable estimates of fixture wattages and HOU, CF, and IF values. A rich body of publicly available research provides these data, which can be found in the resources listed below. Although this is not an exhaustive list, it is representative. Users should select the references that best match their markets and program needs.

The documents cited below have been produced through regulatory and administrative processes, and, as they were developed with considerable oversight and review, they are considered reliable by each sponsoring jurisdiction for their intended applications. HOU, CF, and IF values have been developed from primary data collected during project M&V reviews or evaluation studies, or they are based on engineering analysis. Some of these references provide source documentation.

Fixture wattages are generally based on manufacturers' ratings, obtained during tests conducted according to ANSI standards, although this is not well documented in these sources. Fixture wattages are independent of geographic location. Also, HOU values also tend to be consistent for non-residential building types regardless of location. The sources cited here can be used for these parameters in any service territory.

IF and CF parameters, on the other hand, are dependent on local conditions (weather and system load shape) and users should select carefully so that the referenced values reflect local conditions. Alternatively, local IF and CF parameters can be developed using computer simulations and system load shapes for the service territory where they will be used.

The following documents have informed the development of this protocol. Users will find them a useful starting point in locating the data required to implement the protocol's savings algorithms and procedures.

California Energy Commission (CEC).

DOE *Advanced Lighting Guidelines*.1993

"Database for Energy Efficient Resources (DEER)." California Public Utilities Commission (CPUC).www.deeresources.com.

Efficiency Valuation Organization (EVO). IPMVP, Uncertainty Assessment. Anticipated to be available to the public winter of 2018. Free registration required to download: http://evo-world.org/en/

Massachusetts Program Administrators. (October 2011). *Massachusetts Technical Reference Manual for Estimating Savings from Energy Efficiency Measures* (October 2015). http://ma-eeac.org/wordpress/wp-content/uploads/2016-2018-Plan-1.pdf.

New York Department of Public Service. (2010). *New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs.* Prepared for the New York Department of Public Service. Albany, New York: New York Department of Public Service, pp. 109-270.

Vermont Energy Investment Corporation. (2010). *State of Ohio Energy Efficiency Technical Reference Manual*. Prepared for the Public Utilities Commission of Ohio.

12 Appendix

Table 7. Example Lighting Inventory Form

Project Name:	Example Building Lighting Project #3	**Facility Type:**	Office
Site Name:	Example Building	**Facility Location:**	NYC
Utility Acct Number(s):	XXX-XXXXXXXXXX	**Facility Square Feet**	40,000
Type of Heating Equipment:	Gas fired boiler		
Type of Cooling Equipment:	Rooftop DX	Date(s) Survey completed:	
		Survey completed by (name):	

INSTRUCTIONS

				PRE-INSTALLATION					POST-INSTALLATION										
Area Description	**Usage Group ID**	**Heat?**	**Cool?**	**Pre Fixt. No.**	**Pre Fixt. Code**	**Pre Watts/Fixt**	**Pre kW/Space**	**Existing Control**	**Post Fixt No.**	**Post Fixt Code**	**Post Watts/ Fixt**	**Post kW/Space**	**Proposed Control**	**kW Saved**	**Heating InterActive Factor**	**Cooling InterActive Factor**	**Baseline Annual Hours**	**Proposed Annual Hours**	**Annual kWh Saved**
Unique description of the location that matches the site map	Descriptive name for the usage group	Yes or No	Yes or No	Number of fixtures before the retrofit	Code from Table of Standard Fixture Wattages	Value from Table of Standard Fixture Wattages	(Pre Watts/Fixt) * (Pre Fixt No.)	Pre-installation control device	Number of fixtures after the retrofit	Code from Table of Standard Fixture Wattages	Value from Table of Standard Fixture Wattages	(Post Watts/Fixt) * (Post Fixt No.)	Post-installation control device	(Pre kW/Space) - (Post kW/Space)	Change in heating energy due to lighting project	Change in cooling energy due to lighting project	Existing annual hours for the usage group	Propsed annual hours for the usage group	[(Pre kW/Space * Baseline Annual Hours) - (Post kW/Space * Proposed Annual Hours)] * (1+Heat-IF)
Room 343	Office	Yes	Yes	8	2F40SEM	70	0.56	Switch	8	2F25EEE	43	0.34	Switch	0.22	-	0.03	2,500	2,500	558
Room 344	Office	Yes	Yes	3	2F40SEM	70	0.21	Switch	3	2F25EEE	43	0.13		0.08	-	0.03	2,500	2,500	209
Corridor Floor 3	Hallway	Yes	Yes	17	1F40SEE	38	0.65	Switch	17	1F25EEE	30	0.51		0.14	-	0.03	3,700	3,700	520
Women RR Flr 3	Restroom	Yes	Yes	4	110060	60	0.24	Switch	4	1C00185	20	0.08		0.16	-	0.03	3,700	3,700	612
Men RR Flr 3	Restroom	Yes	Yes	4	110060	60	0.24	Switch	4	1C00185	20	0.08		0.16	-	0.03	3,700	3,700	612
	TOTAL			**36.00**		**298.00**	**1.90**		**36.00**	**1.14**	**156.00**	**1.14**		**0.75**					**2,510**

Table 8. New York Standard Approach for Estimating Energy Savings

FIXTURE CODE	LAMP CODE	DESCRIPTION	BALLAST	Lamp/ fix	WATT/ LAMP	WATT/ FIXT
F42SSILL	F28T8	Fluorescent, (2) 48", Super T-8 lamp, Instant Start Ballast, NLO (BF: .85-.95)	Electronic	2	28	48
F41SSILL/T4	F28T8	Fluorescent, (2) 48", Super T-8 lamp, Instant Start Ballast, NLO (BF: .85-.95), Tandem 4 Lamp Ballast	Electronic	2	28	47
F42SSILL-R	F28T8	Fluorescent, (2) 48", Super T-8 lamp, Instant Start Ballast, RLO (BF<0.85)	Electronic	2	28	45
F41SSILL/T4-R	F28T8	Fluorescent, (2) 48", Super T-8 lamp, IS Ballast, RLO (BF<0.85), Tandem 4 Lamp Ballast	Electronic	2	28	44
F42SSILL-H	F28T8	Fluorescent, (2) 48", Super T-8 lamp, Instant Start Ballast, HLO (BF:.96-2.2)	Electronic	2	28	67
F42ILL/T4	F32T8	Fluorescent, (2) 48", T-8 lamp, Instant Start Ballast, NLO (BF: .85-.95), Tandem 4 Lamp Ballast	Electronic	2	32	56
F42ILL/T4-R	F32T8	Fluorescent, (2) 48", T-8 lamp, Instant Start Ballast, RLO (BF<0.85), Tandem 4 Lamp Ballast	Electronic	2	32	51
F42ILL-H	F32T8	Fluorescent, (2) 48", T-8 lamp, Instant Start Ballast, HLO (BF:.96-1.1)	Electronic	2	32	65
F42ILL-R	F32T8	Fluorescent, (2) 48", T-8 lamp, Instant Start Ballast, RLO (BF<0.85)	Electronic	2	32	52
F42ILL-V	F32T8	Fluorescent, (2) 48", T-8 lamp, Instant Start Ballast, VHLO (BF>1.1)	Electronic	2	32	79
F42LE	F32T8	Fluorescent, (2) 48", T-8 lamp	Mag-ES	2	32	71
F42LL	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, NLO (BF: .85-.95)	Electronic	2	32	60
F42LL/T4	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, NLO (BF: .85-.95), Tandem 4 Lamp Ballast	Electronic	2	32	59
F42LL/T4-R	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, RLO (BF<0.85), Tandem 4 Lamp Ballast	Electronic	2	32	53
F42LL-H	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, HLO (BF:.96-1.1)	Electronic	2	32	70
F42LL-R	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, RLO (BF<0.85)	Electronic	2	32	54
F42LL-V	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, VHLO (BF>1.1)	Electronic	2	32	85
F42SE	F40T12	Fluorescent, (2) 48", STD lamp	Mag-ES	2	40	86
F42GHL	F48T5/HO	Fluorescent, (2) 48", STD HO T5 lamp	Electronic	2	54	117
F42SHS	F48T12/HO	Fluorescent, (2) 48", STD HO lamp	Mag-STD	2	60	145
F42SIL	F48T12	Fluorescent, (2) 48", STD IS lamp, Electronic ballast	Electronic	2	39	74
F42SIS	F48T12	Fluorescent, (2) 48", STD IS lamp	Mag-STD	2	39	103

(New York Department of Public Service 2010)

Table 9. New York Standard Approach for Estimating Energy Savings

Facility Type	Lighting Hours	Facility Type	Lighting Hours
Auto Related	4,056	Manufacturing Facility	2,857
Bakery	2,854	Medical Offices	3,748
Banks	3,748	Motion Picture Theatre	1,954
Church	1,955	Multi-Family (Common Areas)	7,665
College – Cafeteria (1)	2,713	Museum	3,748
College - Classes/Administrative	2,586	Nursing Homes	5,840
College - Dormitory	3,066	Office (General Office Types) (1)	3,100
Commercial Condos (2)	3,100	Office/Retail	3,748
Convenience Stores	6,376	Parking Garages	4,368
Convention Center	1,954	Parking Lots	4,100
Court House	3,748	Penitentiary	5,477
Dining: Bar Lounge/Leisure	4,182	Performing Arts Theatre	2,586
Dining: Cafeteria / Fast Food	6,456	Police / Fire Stations (24 Hr)	7,665
Dining: Family	4,182	Post Office	3,748
Entertainment	1,952	Pump Stations	1,949
Exercise Center	5,836	Refrigerated Warehouse	2,602
Fast Food Restaurants	6,376	Religious Building	1,955
Fire Station (Unmanned)	1,953	Restaurants	4,182
Food Stores	4,055	Retail	4,057
Gymnasium	2,586	School / University	2,187
Hospitals	7,674	Schools (Jr./Sr. High)	2,187
Hospitals / Health Care	7,666	Schools (Preschool/Elementary)	2,187
Industrial - 1 Shift	2,857	Schools (Technical/Vocational)	2,187
Industrial - 2 Shift	4,730	Small Services	3,750
Industrial - 3 Shift	6,631	Sports Arena	1,954
Laundromats	4,056	Town Hall	3,748
Library	3,748	Transportation	6,456
Light Manufacturers (1)	2,613	Warehouse (Not Refrigerated)	2,602
Lodging (Hotels/Motels)	3,064	Waste Water Treatment Plant	6,631
Mall Concourse	4,833	Workshop	3,750

(New York Department of Public Service 2010)

Table 10. Midstream Baseline Wattage, Linear Lamps, and Fixtures; HID Interior and Exterior Fixtures (Pennsylvania Public Utility Commission 2016)

Efficient Lamp or Fixture	Minimum Lumen	Maximum Lumen	Watts Base	Note
Highbay & Lowbay LED Fixture	3850	6550	189	150-watt HID lamp
	6551	9300	215	175-watt HID lamp
	9301	11150	241	200-watt HID lamp
	11151	12200	295	250-watt HID lamp
	12201	15550	365	320-watt HID lamp
	15551	20100	462	400-watt HID lamp
	20101	34700	843	750-watt HID lamp
	34701	57250	1090	1000-watt HID lamp
Exterior Fixture (Pole, Wall Pack or Parking Garage)	250	4650	133	100-watt HID lamp
	4651	7900	215	175-watt HID lamp
	7901	11050	295	250-watt HID lamp
	11051	24700	462	400-watt HID lamp
	24701	40750	843	750-watt HID lamp
	40751	54650	1090	1,000-watt HID lamp

Chapter 3: Commercial and Industrial Lighting Controls Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

Stephen Carlson
DNV-GL
Madison, Wisconsin

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68559
September 2017

Contract No. DE-AC36-08GO28308

Chapter 3: Commercial and Industrial Lighting Controls Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

Stephen Carlson
DNV-GL
Madison, Wisconsin

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68559
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Suggested Citation

Carlson, Stephen. (2017). *Chapter 3: Commercial and Industrial Lighting Controls Evaluation Protocol*, *The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68559. http://www.nrel.gov/docs/fy17osti/68559.pdf

Acronyms

BMS	building management system
CEE	Consortium for Energy Efficiency
CSF	control savings factor
CT	current transformer
DOE	U.S. Department of Energy
EFLH	equivalent full load hours
HID	high-intensity discharge
HVAC	heating, ventilating, and air conditioning
IF	interactive factor
IPMVP	International Performance Measurement and Verification Protocol
kW	kilowatt
kWh	kilowatt-hour
M&V	measurement and verification
NYSERDA	New York State Energy Research and Development Authority
OS	occupancy sensors
UMP	Uniform Methods Project

Table of Contents

List of Tables

1 Measure Description

This Commercial and Industrial Lighting Controls Evaluation Protocol (the protocol) describes methods to account for energy savings resulting from programmatic installation of lighting control equipment in large populations of commercial, industrial, government, institutional, and other nonresidential facilities. This protocol does not address savings resulting from changes in codes and standards, or from education and training activities.[1] When lighting controls are installed in conjunction with a lighting retrofit project, the lighting control savings must be calculated parametrically with the lighting retrofit project so savings are not double counted.[2]

An "energy efficiency measure" can be defined as a set of actions and equipment changes—compared to standard or existing practices—resulting in reduced energy use, while maintaining the same or improved service levels for customers or processes.

In addition to delivering light levels required for activities or processes in facilities, lighting control measures shut off lighting fixtures when a space is unoccupied, or operate lighting at reduced power when ambient light levels are high. For retrofit installations, the baseline condition typically equals the lighting operating at normal power levels or when the space is both occupied and unoccupied during normal business hours.[3] New construction baseline conditions are generally provided by state and local building codes. Although codes vary widely throughout the country, typically they require some form of control for most interior lighting. This document includes a detailed discussion of baselines.

Lighting control measures in commercial, industrial, and other nonresidential facilities include:

- Sweep controls/energy management systems that shut off lighting at a set time, typically after normal operating hours
- Lighting occupancy sensors (OS) that turn lights on or off, based on space occupancy conditions
- Dimming control systems:
  - Stepped dimming systems, such as dual ballasts (inboard/outboard)
  - Dual ballast high/low high-intensity discharge (HID)[4]
  - Continuous daylight dimming systems.

---

[1] This protocol addresses only automated lighting control measures, which do not require behavioral actions by space occupants (such as "tuning" light levels for different tasks).

[2] Typically, post-lighting retrofit wattages are used to calculate the lighting controlled kilowatt (kW) value for lighting control savings calculations.

[3] In this case "normal" refers to fixtures operating at full power, and is applicable for all forms of lighting control applications during business operating hours.

[4] Such HID fixtures typically have only one lamp that can be operated at two different output levels by a two stage ballast; this differs from stepped dimming systems that dim by controlling lamps powered by a single ballast.

2 Application Conditions of Protocol

Historically, lighting control equipment has accounted for a relatively small portion of cost-effective, electric energy efficiency resources in the United States. However, use of lighting controls has been increasing due to building efficiency certification standards (such as Leadership in Energy and Environmental Design) and the increased prevalence of demand-response programs.

Typically, lighting controls do not provide a sufficiently large component of an energy efficiency program to warrant their own evaluation efforts, so these measures tend to be included as small parts of commercial and industrial program evaluation. Thus, little effort has been expended to move beyond post-installation metering or applying a 30% control savings factor (CSF).[5]

This evaluation protocol assumes a focus on lighting controls, and that primary data captured will be used to inform the evaluation, or to determine deemed savings in a technical reference manual. By following the methods presented here, evaluators can determine energy savings resulting from lighting controls installed through efficiency programs in a manner that is consistent across jurisdictions or regions. Resulting data will provide planners, policymakers, regulators, and others with sound, comparable information for use in comprehensive energy planning.[6]

The protocol applies to installation of commercial, industrial, and nonresidential lighting control measures in customer facilities; installations result from energy efficiency programs, which have varying delivery methods, depending on target markets and customer types. Primarily, the delivery method can be distinguished by parties receiving incentive payments from a program. Although methods described in this protocol apply to all programs, issues with customer and baseline equipment data vary with each. Common program delivery types include:

1. **Incentive and Rebate**: Under this delivery method, administrators pay program participants in target markets for installing lighting control measures. Participants receive an incentive payment, based on annual energy savings ($/kilowatt-hour [kWh]) for each installed measure, or based on demand savings ($/kW). Participants include design teams, contractors, building owners, and building managers. Savings can be estimated through one or more of the following techniques:

    - Simple engineering calculations
    - A measurement and verification (M&V) process that measures key parameters, such as equivalent full load hours (EFLH), controlled fixture wattages, or a CSF as part of project implementation.

---

5 The 30% savings percentage for OS has been adopted and borrowed in so many technical reference manuals and public savings documents that its exact origin is difficult to trace. Table 4 in this document is an ASHRAE table of control savings factors, and the values range from 0.10 to 0.40 depending on the type of control.

6 As discussed in *Considering Resource Constraints* in the "Introduction" of this UMP report, small utilities (as defined under the Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

Programs also may pay rebates for specific lighting control equipment types (for example, ceiling-mounted OS), with the program using assumptions about "replaced equipment." Thus, increased administrative efficiency is exchanged for less certainty about baseline conditions and, therefore, savings. This type of program implementation approach is often referred to as a "deemed" savings approach where savings are developed on a per unit basis and very little site-specific information is required to determine the claimed (*ex ante)* program savings estimate.

Incentive programs often collect more detailed baseline data than do rebate programs. This includes extensive data about controlled equipment wattages and hours of operation, which facilitates determination of savings impacts, typically using a savings calculation based on these site-specific data. Although rebate programs typically begin with useful information regarding the quantity of lighting control equipment, these programs do not always collect data about controlled fixtures, because it is not necessary to calculate the claimed program savings.

2. **Direct Install**: Using this delivery method, contractors engaged through a program install lighting control equipment in customer facilities. The programs pay contractors directly, while customers receive a lighting control measure free or at a reduced cost. Direct-install programs target hard-to-reach customers—typically small businesses—overlooked by contractors working through incentive and rebate programs. Direct-install programs typically do not focus on lighting control measures, but they may be eligible measures.

In addition to their distinctive delivery methods, commercial, industrial, and nonresidential lighting programs (which include lighting controls) can be classified as targeting retrofit (serving existing facilities) or new construction markets. The program delivery types described above apply to existing building programs. New construction programs primarily employ incentives and rebates to acquire energy efficiency reductions.

New construction programs present evaluators with a dilemma in establishing baselines for buildings that previously did not exist. This problem can be addressed by referring to new construction energy codes for commercial, industrial, and nonresidential facilities (usually by referencing ASHRAE Standard 90.1 or the International Energy Conservation Code). The ASHRAE Standard defines lighting controls under section 9.4.1; these are mandatory for interior lighting in buildings larger than 5,000 ft^2.[7] Other federal, state, and local standards may establish additional baseline constraints on lighting controls.

[7] ASHRAE 90.1, 2004, page 61 addresses mandatory provisions and exceptions for lighting controls in newly constructed buildings.

3 Savings Calculations

Project and program savings for lighting controls and other technologies result from the difference between retrofit use and use that would have occurred had the measure not been implemented (the baseline). The fundamental savings equation is:

> Energy or Demand Savings = (Baseline Period Energy Use – Reporting-Period Energy Use) ± Adjustments

The equation's adjustment term calibrates baseline and/or reporting use and demand to the same set of conditions. Common adjustments account for changes in schedules, occupancy rates, weather, or other parameters that shift between baseline and reporting periods. Adjustments commonly are applied to heating, ventilating, and air-conditioning (HVAC) measures, but less commonly to lighting measures (or adjustments are inherent in algorithms for calculating savings).

3.1 Algorithms

The following equations calculate primary energy savings for lighting control measures in commercial, industrial, and nonresidential facilities.

Equation 1: Lighting Control Electric Energy Savings

$$\text{kWh Save}_{lc} = \text{kW}_{controlled} * \text{EFLH}_{pre} * \text{CSF}$$

where:

kWh Save_{lc} = Annual kWh savings resulting from the lighting control project

$\text{kW}_{controlled}$ = Sum (Fixture Wattage * Quantity Fixtures) for controlled fixtures

EFLH_{pre} = Annual equivalent full load hours prior to application of controls

CSF = Control savings factor is the annualized reduction factor calculated across the EFLH

Equation 1A: Lighting Control Savings Factor

$$\text{CSF} = 1 - (\text{EFLH}_{post} / \text{EFLH}_{pre})$$

where:

CSF = Control savings factor is the annualized reduction factor calculated across the EFLH

EFLH_{pre} = Annual equivalent full load hours prior to application of controls

EFLH_{post} = Annual equivalent full load hours after application of controls

When calculating the site level CSF using measured data for multiple control points, the weighted average should be developed by using the kW controlled as the weighting factor.

Equation 2: Interactive Cooling Electric Energy Savings

$$kWh_{interact - cool} = kW_{cool} \times IF_c \times Hours_{cool}$$

Equation 3: Interactive Heating Electric Energy Savings

$$kWh_{interact - heat} = kW_{heat} \times IF_h \times Hours_{heat}$$

where:

$kWh_{interact - cool}$ = Interactive cooling savings from the lighting control project

kW_{cool} = Mean kW reduction coincident with the cooling hours

$Hours_{cool}$ = Hours when the space is in cooling mode

IF_c = Interactive cooling factor, ratio of cooling energy reduction per unit of lighting energy; caused by reductions in lighting waste heat removed by an HVAC system

$kWh_{interact - heat}$ = Interactive heating savings due to lighting control project: a negative value

kW_{heat} = Mean kW reduction coincident with the heating hours

$Hours_{heat}$ = Hours when the space is in heating mode

IF_h = Interactive heating factor, ratio of heating energy increase per unit of lighting energy; caused by reductions in lighting heat removed by an HVAC system

Equation 4: Total annual energy savings

$$kWh\ Save_{total} = kWh\ Save_{lc} + kWh_{interact - cool} + kWh_{interact - heat}$$

4 Role of the Lighting Control Program Implementer

Successful application of the protocol requires standard data, collected in a prescribed format, as part of the implementation process. The protocol also requires tracking project and program savings estimated on the basis of the standard data.

The implementer is responsible for ensuring collection of data required to track program activity and calculate savings at the project level. The implementer also is responsible for maintaining a program activity record, including anticipated savings by project.

4.1 Implementation Data Requirements

The protocol recommends that, for all projects, the program implementer collect and archive data needed to execute the savings algorithms. These data include:

- Controlled fixture inventory, including fixture wattage
- Controlled fixture quantities
- Controlled fixture baseline lighting EFLH
- Control savings factor
- Usage group assignments
- Heating and cooling equipment types
- Interactive factor for cooling (optional)
- Interactive factor for heating (optional).

Facilities (or spaces within facilities where the project has been installed) are classified as cooled/uncooled and heated/unheated, and information about heating and cooling equipment and fuel types for each should be recorded. This information is required to estimate interactive effects.

4.2 Implementation Data Collection Method

The protocol recommends participants collect and submit required data as a condition for program enrollment. The protocol also recommends the implementer specify data reporting formats, either by supplying a structured form (such as a spreadsheet), or by specifying data fields and types used when submitting material to the program. The format must be electronic, searchable, and sortable, and must support combining multiple files into single tables for analysis by the implementer. Faxes, PDFs, and JPEG formats do not meet these criteria. Microsoft Excel and comma-separated text files are acceptable formats.

The data reporting format should be structured to allow verification of project installations. Each record or line in the report represents a collection of identical fixture types, installed in an easily located room, floor, or space, and belonging to one usage group. Table 1 lists fields required in the data reporting format.[8]

[8] The data sources for these fields are described in section 6.5 *Data Requirements and Sources* of this protocol.

Table 1. Required Lighting Control Data Fields

Field	Note
Location	Floor number, room number, and other descriptions
Usage group	
Location cooling	Yes/no
Conditioned floor area	Square footage of conditioned space
Location cooling type	Water cooled chiller, air cooled chiller, packaged DX, etc.
Location cooling fuel	Electric, non-electric
Location heating	Yes/no
Location heating type	Boiler steam/hydronic, heat pump, forced air, strip heat, etc.
Location heating fuel	Electric, non-electric
Controlled fixture type	From lookup table supplied by implementer, manufacturers cut sheet
Controlled fixture count	
Controlled fixture wattage	From lookup table supplied by implementer, manufacturers cut sheet
Pre-control EFLH	Requirement for pre-metering depends on control type
CSF	Will be calculated using pre/post or post only data
IF_c	Interactive factor for cooling, from lookup table, optional
IF_h	Interactive factor for heating, from lookup table, optional
kWh_{save}	Will be calculated using pre/post or post only data
Measure Cost	Cost of measure in dollars
Incentive Cost	Cost of incentive in dollars

For each project, lighting contractors or other program participants should record:

- Types, quantities, and wattages of all lamps, ballasts and fixture types controlled by a lighting control measure
- All lighting control equipment types and locations throughout the facility.

For lighting control measures reducing power outputs of fixtures, dimming controls must also be described so each increment of light reduction can have an appropriate kW value established. Daylight dimming systems should have ambient light sensor locations identified and minimum power levels specified so the system can be modeled using building simulation software, if necessary. (Sensor location is not required if using a spreadsheet savings estimation approach.)

The protocol recommends integrating savings verification into the program administrative process, particularly for data tracking. Impact evaluations of lighting efficiency and lighting controls programs remain highly dependent on data developed in conjunction with the lighting retrofit construction process. These data should be collected and reported by the project contractor.

The administrator should employ a third-party expert to conduct periodic, systematic reviews and inspections of a sample of completed projects to verify the accuracy of data from the lighting inventory forms. At first, the sampling procedure should be implemented randomly on an approximately 10% fixed-percentage basis so the contractor cannot predict projects to be inspected. In addition to requiring the contractor to correct discrepancies, the administrator may choose to impose penalties for egregious or repeated errors. Once a contractor has proven reliable, the sampling percentage can be reduced, but the random sampling procedure should be maintained.

5 Role of the Evaluator

The evaluator's role is to determine energy savings resulting from the operation of lighting control efficiency programs. The procedure reviews a sample of completed projects, including conducting on-site M&V activities, calculating a realization rate (the ratio of evaluator to implementer anticipated savings), and using the realization rate to adjust implementer-anticipated savings.

5.1 Evaluator Data Requirements

The protocol recommends the program evaluator collect the same data as the implementer. As described in M&V, the evaluator must have access to implementation lighting inventory forms and participant application materials for each project in the sample.

5.2 Evaluator Data Collection Method

Under the protocol, the implementer provides the evaluator with a copy of the program and project data tracking record for the evaluation review period. This record includes the fields shown in Table 1. The implementer also provides all records for projects in the evaluation review sample, including application materials and site contact information.

The protocol recommends the evaluator collect additional M&V data during site visits conducted for the sample of evaluation review projects.

Table 2 lists the data required for each project in the evaluation sample.

Table 2. Lighting Control Data Required by Evaluator

Field	Note
Location	From implementer
Usage group	From implementer
Location cooling	From implementer, verified by evaluator
Location cooling type	From implementer, verified by evaluator
Location cooling fuel	From implementer, verified by evaluator
Location heating	From implementer, verified by evaluator
Location heating type	From implementer, verified by evaluator
Location heating fuel	From implementer, verified by evaluator
Controlled fixture type	From implementer, verified by evaluator
Controlled fixture count	From implementer, verified by evaluator
Controlled fixture wattage	From implementer, verified by evaluator
Pre-control EFLH	From implementer, could be measured by evaluator
CSF	Measured by evaluator
IF_c	Interactive factor for cooling, from lookup table, optional
IF_h	Interactive factor for heating, from lookup table, optional
kWh_{save}	Will be calculated using pre/post or post only data

6 Measurement and Verification Plan

The M&V plan describes how evaluators determine verified energy savings in a facility where a lighting controls efficiency project has been installed. M&V results are applied to the population of all completed projects to determine program savings. All M&V activities in the protocol are conducted for a representative sample of completed projects. The evaluator is responsible for meeting M&V requirements in the protocol.

6.1 IPMVP Option

The selection of the appropriate International Performance Measurement and Verification Protocol (IPMVP) evaluation method for reporting evaluated (*ex post*) savings is contingent on site-specific criteria. The key factors for determining the method are the availability of whole premise interval metered data and preferably sub-metered lighting data, and the relative size of the savings impact attributable to the lighting control measure. When the savings impact for the lighting control measure is at least 5%, and preferably at least 10% of the energy usage for the available interval data, then IPMVP Option C–Whole Facility should be selected.[9] When Option C is selected, there must be both pre- and post-metered data available to evaluate the lighting control impacts. Because lighting controls often do not meet the relative impact criteria, the IPMVP Option A–Retrofit Isolation: Key Parameter Measurement approach is the most common method used to evaluate savings. Key parameters to be measured include $EFLH_{pre}$ and $EFLH_{post}$, to calculate the CSF term in Equation 1. Accurately measuring these variables typically requires determining lighting usage in the pre-control state, and may require measuring usage in the post-control state.[10]

Table 3 provides metering recommendations for measuring various types of lighting control measures. In summary:

- Lighting sweep controls, energy management systems, and time clock measures require pre- and post-installation metering to establish EFLH and CSF accurately.
- OS measures can be determined effectively through pre-installation metering only if using a lighting event logger with infrared occupancy sensor capabilities.[11]
- Most dimming applications can be measured using post-installation data only when these are sufficiently accurate to assume uncontrolled kW would equal controlled lighting operating at full power.
- Event loggers typically are lighting loggers monitoring lighting on/off operations via a photocell; power loggers monitor power consumption of controlled lighting circuits.

[9] In this case, the data could be either whole premise data or lighting end use data, which contain the savings attributable to the lighting control measure(s) as a portion of the data. In either case the savings impact must be at least 5% of the total usage observed in the data in order to quantify the impacts accurately using this method.

[10] IPMVP Option A - Retrofit Isolation requires the key savings variable be measured pre and post. However, when conducting M&V in an impact evaluation, it can be a challenge to obtain baseline data. The program administrator often does not collect the data, and evaluators commonly do not become involved until after the project is installed.

[11] These loggers monitor lighting on/off as well as whether the space is occupied or unoccupied. These data, coupled with the lighting latency factor, can be used to establish EFLH and CSF. Some companies maintain these data by building type and space, offering data that can be purchased: www.sensorswitch.com/Literature.aspx

Table 3. Metering Requirements for Various Lighting Control Strategies

Lighting Control Measure	Metering Recommendations: Pre-Installation	Metering Recommendations: Post-Installation	Metering Type
Lighting Sweep Controls/Energy Management System/Time Clock	Yes	Yes	Event or Power Logger
Occupancy Sensors	Yes	Yes/No	Event/Event and Occupancy Logger
Stepped Dimming (Dual Ballasts)	No	Yes	Event Logger
Dual Ballast (High/Low HID)	No	Yes	Power Logger
Continuous Daylight Dimming	No	Yes	Power Logger

Additionally, ASHRAE recommends that lighting levels be measured for lighting control measures—particularly dimming measures—to make sure that adequate lighting levels at the work area are maintained.

6.2 Verification Process

Verification involves visual inspections and engineering calculations to establish an energy efficiency project's potential to achieve savings. The verification process determines the controlled fixture wattage and controlled fixture quantity parameters used to calculate the $kW_{controlled}$ variable in Equation 1. The following describes activities involved in the process:

1. Select a representative sample of projects for review. (See Chapter 11: *Sample Design Protocol* for guidance on sampling.)
2. Schedule a site visit with a facility representative for each project in the sample.
3. Conduct an on-site review for each project. Inspect a representative sample of controlled lighting fixtures and lighting controls reported by the implementer and verify that the controls are operating as reported. (See the "Sample Design" protocol for guidance on sampling.)
    a. Confirm or correct reported controlled fixture types and wattages for each fixture in the sample.
    b. Confirm or correct reported quantities for all controlled fixtures in the sample.
    c. Confirm or correct the heating/cooling status and associated equipment for spaces in the sample.
    d. Interview facility representatives to check baseline fixture control types and quantities reported for the sample. Confirmation or correction will be based on the interviews. When available, interviews are supplemented by physical evidence such as lighting controls installed on fixture types or in areas not changed by the project.
4. Update the lighting control inventory form for the sample, based on findings from the on-site review.

At completion of the verification process, the evaluator will have confirmed or corrected fixture wattage and fixture quantity parameters used to calculate the $kW_{controlled}$ variable in Equation 1.

6.3 Measurement Process

The measurement process involves using electronic metering equipment to collect data determining EFLH and CSF parameters in Equation 1. Usually, equipment is installed temporarily during the measurement period; in some facilities, existing building automation systems monitoring lighting circuits may be employed. Lighting control measures particularly can be challenging to measure as they may require use of pre/post metering of either on/off operations or interval power consumption.

Meters and metering data used to measure lighting control operating characteristics either record a change of state (light on, light off), or continuously sample and record current or power on a lighting circuit or reduced light output of a fixture. All data must be time-stamped for application in the protocol.

Temporary metering equipment, in the form of data loggers, is most commonly used for establishing lighting EFLH.

Change-of-state lighting data loggers are small (matchbox sized), integrated devices that include a photocell, microprocessor, and memory. These data loggers are mounted inside fixtures. Each time lamps in the fixtures are turned on or off, the event is recorded and time stamped. Such lighting loggers are only suitable for monitoring on/off lighting controls, such as OS, lighting sweep controls/energy management systems, and stepped dimming systems (for example, inboard/outboard configurations, where controlled lamps can be isolated from uncontrolled lamps). For lighting control systems that vary lighting power, such as dimming systems or dual ballast HID systems in which the lamps cannot be isolated, interval power of the lighting system must be monitored.

Data loggers continuously sampling and recording lighting operating hours information usually require an external sensor, such as a current transformer (CT) or photocell. Data loggers with CTs can monitor amperage to a lighting circuit. Spot measurements of the circuit's amperage with lights on and off establishes threshold amperages for on conditions. Similarly, data loggers with an external photocell can record light levels in a space. Spot measurements of lumen levels with the fixtures on and off establishes light level thresholds for on conditions. Data loggers are commonly used for amperage measurement; continuous light level monitoring to determine hours of operation is less common.

Data logger failures due to incorrect adjustments, locations, or software launches occur commonly. The protocol recommends carefully following manufacturer's recommendations.

Measurement involves metering lighting operating hours for a representative sample of controlled fixtures selected for verification. Meters are deployed (or metering routines are established, if using an existing building management system [BMS]) during the verification site visit. The process requires the following activities:

1. Meter operating hours for each circuit in the verification sample.
    a. If using light loggers, deploy loggers in one or more fixtures controlled by the circuit. Only one logger per last point of control is required; however, additional loggers are commonly deployed to offset logger failure or loss.

b. If measuring amperage, install the CT and data logger in lighting panels for the sampled circuit. The sampling interval should be 15 minutes or less. Spot-measure amperage with lights on and off for the circuit leg with the CT. Record the amperage threshold for the lights-on condition.

c. If the lighting control measure is an on/off type of control (such as occupancy sensors), an event type power logger can be used. Event power loggers record a change of state when the power is on and off and provide similar data as a change of state lighting logger. The sampling interval is irrelevant for event loggers because it captures transitions and data can be output at any interval desired.

d. If using a BMS, establish trends for lighting on/off status for each circuit in the sample. The sampling interval should be 15 minutes or less. Check that the BMS has sufficient capacity to archive recorded data, and that the metering task will not adversely slow the BMS response time.

2. Check data logger operations. Before leaving the site, spot-check a few data loggers to confirm they are recording data as expected. Correct any deficiencies, and, if they appear systemic, redeploy the loggers. If using BMS trends, spot-check recorded data.
3. Leave metering equipment for the monitoring period, which could include pre and post periods. The protocol recommends a monitoring period capturing the full range of facility operating schedules. For facilities with constant schedules (such as office buildings, grocery stores, and retail shops), the protocol calls for metering a minimum of two weeks for pre periods and a minimum of four weeks for post periods. Facilities with variable schedules will require additional time. Facilities with seasonal schedules, such as schools, should be monitored during active periods.
4. Analyze metering data. Calculate the percent-on time for metered lighting equipment for each usage group. When pre-control data are collected for control systems, pre-control EFLH can be calculated directly, and post EFLH can be calculated as well. In this case, the CSF equals 1 minus the ratio of post EFLH, divided by pre EFLH. For lighting control measures varying seasonally, such as continuous daylight dimming systems, it will be necessary to annualize metered data to account for daylight hours during the metering period so summer metering does not over-predict savings, or winter metering does not under-predict savings. Similarly, facilities with seasonal schedules, such as schools, which should have been metered during active periods, have annual EFLH and CSF values adjusted to reflect operating schedules.

6.4 Report M&V Savings

Information collected during the M&V processes can be used to calculate M&V project savings as follows:

1. Using results from the last step in the measurement process and the sample lighting inventory form from the verification process, update the inventory EFLH and CSF parameters and calculate M&V savings for the sample.
2. Calculate the project realization rate: the ratio of M&V savings to savings reported by the implementer for the sample.

3. Calculate project M&V savings: the product of the project realization rate and project savings reported by the implementer.
4. Site level savings estimates are used to develop program level results and are weighted and expanded based upon the sample design to develop program level realization rates and statistical relative precision at the selected confidence interval.[12]

6.5 Data Requirements and Sources

Data requirements are described in *Role of the Lighting Control Program Implementer* and *Role of the Evaluator*, with additional detail included in the M&V. This section addresses information on the fixture wattage, EFLH, and CSF parameters in the algorithm equations.

6.5.1 Fixture Wattage

The protocol recommends use of fixture wattage tables, developed and maintained by existing energy efficiency programs and associated regulatory agencies. The tables list all common fixture types, and most are updated as new fixtures and lighting technologies become available. Wattage values are measured according to American National Standards Institute standards[13] by research facilities working on behalf of manufacturers and academic laboratories.

In the wattage table, each fixture and screw-in bulb is fully described, and assigned a unique identifier. The implementer enters a fixture code into the lighting inventory form, which automatically performs a lookup function to enter the associated demand into the form. The evaluator verifies or corrects the fixture type for the evaluation sample in a copy of the implementer's inventory form, automatically updating lighting values.

The protocol recommends adopting a fixture wattage table used by an established and recognized lighting efficiency program. As of May 2012, the following sources serve as examples:

- *Massachusetts Technical Reference Manual 2011*, *Massachusetts Device Codes and Rated Lighting System Wattage Table*. Available from the Massachusetts Energy Efficiency Advisory Council: www.ma-eeac.org/index.htm. This is a slightly abbreviated and simplified table of common fixtures and their wattages.
- *New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs 2010,* Appendix C Standard Fixture Watts. Available from the New York Department of Public Service: www.dps.ny.gov/TechManualNYRevised10-15-10.pdf. This is a comprehensive (34 page) list used by the New York State Energy Research and Development Authority (NYSERDA) since the late 1990s.
- *Database for Energy Efficiency Resources.* Available from the California Public Utilities Commission: www.deeresources.com. An exhaustive list of all parameters

---

12 The confidence interval and testing criteria (one-tail vs. two-tail) can be different based upon jurisdictional requirements. For example, PJM requires relative precision of demand impacts be calculated at 90% confidence using a one-tail test: Independent System Operator-New England requires relative precision of demand impacts be calculated at 80% confidence interval using a two-tail test, both calculations provide the same result.

13 The American National Standards Institute 82.2-2002 test protocol specifies ambient conditions for ballast/lamp combinations in luminaires. The test is conducted on an open, suspended fixture. Actual fixture wattage varies, depending on the installation (suspended, recessed) and housing type. Differences are small, less than 5% (see *DOE 1993 Advanced Lighting Guidelines*.)

driving energy use and savings for a lengthy list of measures. References California codes and weather zones.

An excerpt from the *New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs* is included in the Appendix to this protocol as an example of a wattage table. Wattage tables are used by implementers and evaluators.

6.5.2 EFLH and CSF

EFLH and CSF greatly vary by application. The protocol requires evaluators measure pre- and/or post-EFLH (depending on the control type [see Table 3]) and calculate the CSF to minimize uncertainty.

The following section describes data sources and methods used by program implementers for estimating EFLH and CSF parameters to reliably report project and program savings. The protocol requires program participants to provide estimates of EFLH values by usage group and an estimate of CSF by control type in their lighting inventory forms. The estimate should not be based on the building schedule alone, although this may be used to inform the estimate. The protocol recommends participants develop EFLH and CSF values using one of the following sources, with guidance from the program implementer:

1. Lighting schedules in buildings with energy management systems or time clocks that control lighting equipment. Schedules should be checked by interviewing building managers to determine whether they are overridden. When available, control schedules (or trend data) provide reliable estimates of true lighting operating hours.
2. Interviews with building managers. Building managers are usually familiar with lighting schedules; they may not, however, know how lighting is controlled, and may not be a good source of estimates for CSF values.
3. Tables of EFLH and CSF values by building type, provided by the program implementer.
4. Combinations of interviews and tables.

To calculate and report project savings, the protocol recommends lighting efficiency programs require contractors primarily use deemed EFLH-by-building type values, and use 30% or less for the CSF. When EFLH values can be reliably estimated using site-specific operating schedule data by lighting control usage group, these values should be used to calculate the pre-control EFLH. If the CSF value can be reliably calculated based on the control description, a calculated value should be used *if* the value does not exceed 50% of the published value. Deemed pre-control EFLH and CSF tables should be updated according to a continuous revision schedule so lighting programs using results from logger studies conducted for impact evaluation studies have current information. Table 4 provides a list of lighting CSFs developed from ASHRAE 90.1 power adjustment factors.

Table 4. Lighting Control Savings Factors by Control Type

Lighting Control Type	CSF
Light switch	0
No controls	0
Daylight controls (DC)—continuous dimming	0.3
DC—multiple-step dimming	0.2
DC—ON/OFF	0.1
OS	0.3
OS w/DC—continuous dimming	0.4
OS w/DC—multiple-step dimming	0.35
OS w/DC—ON/OFF	0.35

7 Other Evaluation Issues

7.1 New Construction

Lighting control savings for new construction projects can be difficult to calculate as it can be difficult to monitor pre-controls conditions. In some cases, one may override the controls, as to meter non-control conditions. When possible, EFLH and CSF can be measured using pre/post metering techniques. Overriding controls can also be used for retrofit and incentive programs, providing the site contact is cooperative and the extra site visit is considered in evaluation planning.[14]

7.2 Coincidence Factor

The following equation converts the change in connected load to a demand reduction, coincident with a facility's utility peak period:

Equation 2

$$kW_{reduction,\ coincident} = kW_{controlled} \text{ x } CSF_{coincident}$$

where:

$kW_{controlled}$ = Sum (Fixture Wattage * Quantity Fixtures) for controlled fixtures

$CSF_{coincident}$ = the CSF calculated during the peak period and is equal to the $EFLH_{post}$ during the coincident period divided by the $EFLH_{pre}$ during the coincident period.

IF and CF parameters in Equation 3, Equation 4, and Equation 5 can be measured: (1) determined by measurement, (2) developed from a study of measured data, or (more typically) (3) deemed based on prior studies and computer simulations. Resources for IF and CF values are provided at the end of this document.

[14] New Construction baselines may be irrelevant as lighting controls have mandatory provisions in recent standards, such as ASHRAE Standard 90.1-2004, requiring some form of lighting controls. For programmatic savings, any controls must exceed minimum baseline controls.

8 Program Evaluation Elements

Building a foundation for successful evaluation of a commercial, industrial, and nonresidential lighting controls program begins in the program design phase. Administrators support future evaluations by ensuring the data required to conduct an impact study have been collected, stored, and checked for quality. These data include measured and stipulated values available from prior studies or equipment tests. Administrators must set data requirements before a program's launch so that when data are ultimately reviewed through an impact evaluation, information required to conduct the research will be available.

9 Resources

Note: EFLH, CF, and IF values as well as individual fixture wattages can be found in the following references. (The Pennsylvania reference includes an extensive table of fixture wattages.)

American Society of Heating Refrigerating and Air-Conditioning Engineers (ASHRAE). (2004) *ANSI/ASHRAE/IESNA Standard 90.1-2004 Energy Standard for Buildings Except Low-Rise Residential Buildings.*

California Public Utilities Commission (CPUC). (2008). *Database for Energy Efficient Resources (DEER)*. www.deeresources.com.

Federal Energy Management Program (FEMP). (2008). *M&V Guidelines: Measurement and Verification for Federal Energy Projects Version 3.0.* www1.eere.energy.gov/femp/pdfs/mv_guidelines.pdf.

Massachusetts Program Administrators. (2011). *Massachusetts Technical Reference Manual for Estimating Savings from Energy Efficiency Measures 2012 Program Year—Plan Version.* www.masssave.com.

Pennsylvania Public Utility Commission. (2011). *Technical Reference Manual*, Appendix C. www.puc.state.pa.us/electric/Act129/TRM.aspx.

TecMarket Works. (2010). *New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs—Residential, Multi-Family and Commercial/Industrial Measures.* Prepared for the New York Public Service Commission. www.dps.ny.gov/TechManualNYRevised10-15-10.pdf.

Vermont Energy Investment Corporation. (2010). *State of Ohio Energy Efficiency Technical Reference Manual*. Prepared for the Public Utilities Commission of Ohio. http://amppartners.org/pdf/TRM_Appendix_E_2011.pdf.

Federal Energy Management Program (FEMP). (2008). *M&V Guidelines: Measurement and Verification for Federal Energy Projects Version 3.0.* www1.eere.energy.gov/femp/pdfs/mv_guidelines.pdf.

10 Appendix

Table 5. New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs New York Department of Public Service Appendix C: Standard Fixture Watts (excerpt, page 270)

FIXTURE CODE	LAMP CODE	DESCRIPTION	BALLAST	Lamp/ fix	WATT/ LAMP	WATT/ FIXT
F42SSILL	F28T8	Fluorescent, (2) 48", Super T-8 lamp, Instant Start Ballast, NLO (BF: .85-.95)	Electronic	2	28	48
F41SSILL/T4	F28T8	Fluorescent, (2) 48", Super T-8 lamp, Instant Start Ballast, NLO (BF: .85-.95), Tandem 4 Lamp Ballast	Electronic	2	28	47
F42SSILL-R	F28T8	Fluorescent, (2) 48", Super T-8 lamp, Instant Start Ballast, RLO (BF<0.85)	Electronic	2	28	45
F41SSILL/T4-R	F28T8	Fluorescent, (2) 48", Super T-8 lamp, IS Ballast, RLO (BF<0.85), Tandem 4 Lamp Ballast	Electronic	2	28	44
F42SSILL-H	F28T8	Fluorescent, (2) 48", Super T-8 lamp, Instant Start Ballast, HLO (BF:.96-2.2)	Electronic	2	28	67
F42ILL/T4	F32T8	Fluorescent, (2) 48", T-8 lamp, Instant Start Ballast, NLO (BF: .85-.95), Tandem 4 Lamp Ballast	Electronic	2	32	56
F42ILL/T4-R	F32T8	Fluorescent, (2) 48", T-8 lamp, Instant Start Ballast, RLO (BF<0.85), Tandem 4 Lamp Ballast	Electronic	2	32	51
F42ILL-H	F32T8	Fluorescent, (2) 48", T-8 lamp, Instant Start Ballast, HLO (BF:.96-1.1)	Electronic	2	32	65
F42ILL-R	F32T8	Fluorescent, (2) 48", T-8 lamp, Instant Start Ballast, RLO (BF<0.85)	Electronic	2	32	52
F42ILL-V	F32T8	Fluorescent, (2) 48", T-8 lamp, Instant Start Ballast, VHLO (BF>1.1)	Electronic	2	32	79
F42LE	F32T8	Fluorescent, (2) 48", T-8 lamp	Mag-ES	2	32	71
F42LL	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, NLO (BF: .85-.95)	Electronic	2	32	60
F42LL/T4	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, NLO (BF: .85-.95), Tandem 4 Lamp Ballast	Electronic	2	32	59
F42LL/T4-R	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, RLO (BF<0.85), Tandem 4 Lamp Ballast	Electronic	2	32	53
F42LL-H	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, HLO (BF:.96-	Electronic	2	32	70

FIXTURE CODE	LAMP CODE	DESCRIPTION	BALLAST	Lamp/ fix	WATT/ LAMP	WATT/ FIXT
		1.1)				
F42LL-R	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, RLO (BF<0.85)	Electronic	2	32	54
F42LL-V	F32T8	Fluorescent, (2) 48", T-8 lamp, Rapid Start Ballast, VHLO (BF>1.1)	Electronic	2	32	85
F42SE	F40T12	Fluorescent, (2) 48", STD lamp	Mag-ES	2	40	86
F42GHL	F48T5/HO	Fluorescent, (2) 48", STD HO T5 lamp	Electronic	2	54	117
F42SHS	F48T12/HO	Fluorescent, (2) 48", STD HO lamp	Mag-STD	2	60	145
F42SIL	F48T12	Fluorescent, (2) 48", STD IS lamp, Electronic ballast	Electronic	2	39	74
F42SIS	F48T12	Fluorescent, (2) 48", STD IS lamp	Mag-STD	2	39	103

(Reference: NYSERDA Existing Buildings Lighting Table with Circline Additions from CA SPC Table)

NREL
NATIONAL RENEWABLE ENERGY LABORATORY

Chapter 4: Small Commercial and Residential Unitary and Split System HVAC Heating and Cooling Equipment-Efficiency Upgrade Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in April 2013. The content in this version has been updated.

David Jacobson
Jacobson Energy Research
Providence, Rhode Island

Jarred Metoyer
DNV GL
Madison, Wisconsin

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68560
October 2017

Contract No. DE-AC36-08GO28308

Chapter 4: Small Commercial and Residential Unitary and Split System HVAC Heating and Cooling Equipment-Efficiency Upgrade Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in April 2013. The content in this version has been updated.

David Jacobson
Jacobson Energy Research
Providence, Rhode Island

Jarred Metoyer
DNV GL
Madison, Wisconsin

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68560
October 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter authors wish to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Miriam Goldberg of DNV-GL
- Peter Jacobs of Building Metrics
- M. Sami Khawaja and John Walczyk of Cadmus
- Justin Spencer and Frank Stern of Navigant
- Kevin Warren of Warren Energy.

Suggested Citation

Jacobson, D. and Metoyer, J. (2017). *Chapter 4: Small Commercial and Residential Unitary and Split System HVAC Heating and Cooling Equipment-Efficiency Upgrade Evaluation Protocol*. The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68560. http://www.nrel.gov/docs/fy17osti/68560.pdf

Acronyms

AHRI	Air-Conditioning, Heating, and Refrigeration Institute
CEE	Consortium for Energy Efficiency
COPH	heating coefficient of performance
CV	coefficient of variation
ECM	electronically commutated motor
EER	energy efficiency ratio
EFLH	equivalent full-load hour
EM&V	evaluation, measurement, and verification
HSPF	heating seasonal performance factor
HVAC	heating, ventilating, and air conditioning
IEER	integrated energy efficiency ratio
IPLV	Integrated Part Load Value
IPMVP	International Performance Measurement and Verification Protocol
NEEP	Northeast Energy Efficiency Partnerships
NMBE	normalized mean bias error
RMSE	root mean squared error
SEER	seasonal energy efficiency ratio
TRM	technical reference manual
UMP	Uniform Methods Project

Protocol Updates

The original version of this protocol was published in April 2013.

This chapter has been updated to incorporate the following revisions:

- Expanded the protocol to include some heating equipment, including basic upgrades from standard to high efficiency equipment for ductless mini-split heat pumps and air source heat pumps.
- Updated the regression model to include heating, including the incorporation of a change point model for heating and cooling units.
- Updated the example protocols to reflect current programs.
- Revised data requirements to include more detailed model numbers with imbedded information.

Table of Contents

List of Tables

1 Measure Description

A packaged system—often called a “rooftop unit” because it is usually installed on the roof of a small commercial building—puts all cooling and ventilation system components (evaporator, compressor, condenser, and air handler) in one enclosure or package. The capacity of packaged systems typically ranges from 3 to 20 tons, although specifications go up to 63.3 tons.

Split systems primarily are used for residences and very small commercial spaces. These systems place condensers and compressors outdoors and place evaporators and supply fans indoors. On average, split systems have a capacity of less than 65,000 Btu/hr (5.4 tons).[1] Small systems are rated using the Air-Conditioning, Heating, and Refrigeration Institute (AHRI) standard 210/240, while the large systems are rated using AHRI 340/360. For this protocol, split systems include ductless mini-split heat pumps and air source heat pumps. We recommend applying the protocol to applications where the unit serves a single conditioned zone. We recommend the enhanced or other more advanced methods for larger multi-zone units connected to a variable air volume system.

[1] A ton equals 12,000 Btu/hr, or the amount of power required to melt 1 ton of ice in 24 hours.

2 Application Conditions of Protocol

The specific measure described here involves improving the overall efficiency in air-conditioning systems as a whole (compressor, evaporator, condenser, and supply fan). The efficiency rating for cooling is expressed as the energy efficiency ratio (EER), seasonal energy efficiency ratio (SEER), and integrated energy efficiency ratio (IEER). The rated heating efficiency (applicable for heat pumps only) is expressed as either the heating seasonal performance factor (HSPF) or the heating coefficient of performance (COPH). The higher the EER, SEER, or IEER, the more efficient the unit is. The same applies for the rated HSPF and COPH.

- EER is the Btu/hr of peak cooling delivered per watt of electricity used to produce that amount of cooling. Generally, the EER is measured at standard conditions (95°F outdoor dry bulb, 67°F indoor wet bulb), as determined by the AHRI Standard 210/240 (AHRI 2008 with 2012 Addenda).
- SEER is a measure of a cooling system's efficiency during the entire cooling season for units with rated cooling capacities of less than 65,000 Btu/hr (less than 5.4 tons). The AHRI 210/240-2008 standard describes the tests and calculation method to determine SEER.[2] The SEER is also expressed in Btu/hr cooling per watt of electric input.
- IEER is a measure of a cooling system's efficiency during the entire cooling season for units with cooling capacities between 65,000 Btu/hr (5.4 tons) and 760,000 Btu/hr (63.3 ton), expressed in Btu/hr of cooling per watt of electric input. AHRI Standard 340/360 2015 defines IEER as "a single number figure of merit expressing cooling part load EER for commercial unitary air-conditioning equipment and heat pump equipment on the basis of weighted operation at various load capacities." It replaces the Integrated Part Load Value (IPLV) in ASHRAE standard 90.1-2007 (AHRI 2015).
- HSPF is a measure of a heat pump system's efficiency during the entire heating season for units with rated capacities of less than 65,000 Btu/hr. The AHRI 2010/240-2008 standard descries the tests and calculation method to determine the HSPF.[3]
- COPH is a ratio of the heating capacity in watts to the power input in watts at any given set of rating conditions, expressed in W/W. For COPH, supplementary resistance is excluded. The high and low temperature COPH are defined at the following conditions (AHRI Standard 340/360 2015):
    - High Temperature Coefficient of Performance, COPH, W/W, at 47°F
    - Low Temperature Coefficient of Performance, COPH, W/W, at 17°F.

For many commercial unitary rebate programs offered in 2013 through 2015, units greater than 5.4 tons qualified based on the EER only, and IEER is not captured. Although IEER provides a

[2] SEER was designed to represent the cooling seasonal efficiency in an average U.S. climate. The seasonal efficiency of a unit may vary across climates (Fairey et al. 2004).
[3] HSPF was designed to represent the heating seasonal efficiency in an average U.S. climate. The seasonal efficiency of a unit may vary across climates (Fairey et al. 2004).

more accurate measure of seasonal efficiency for larger units, its use is not yet commonplace throughout the incentive program community.

Table 1 presents a typical program offering for this measure. There are similar programs with additional references from Wisconsin and California.[4] The Consortium for Energy Efficiency (CEE) (2016) continues to provide consistent efficiency "tiers" across the country.

Table 1. Typical Incentive Offering for Air-Cooled Unitary AC, Split Systems, and Heat Pumps (New Condenser and New Coil)

Minimum Efficiency Levels/Incentive Levels						
HVAC Unit Size			Level 1		Level 2	
Tons	Btuh		Minimum SEER/EER for Incentive	Incentive $/Ton	Minimum SEER/EER for Incentive	Incentive $/Ton
Air Conditioning Systems						
Air Cooled Unitary and Split Air Conditioning Systems						
< 5.4	< 65,000	Split	14.0 SEER & 12.0 EER	$70	15.0 SEER & 12.5 EER	$125
< 5.4	< 65,000	Packaged	14.0 SEER & 11.6 EER	$70	15.0 SEER & 12.0 EER	$125
≥ 5.4 to < 11.25	≥ 65,000 to < 135,000		11.5 EER	$50	12.0 EER	$95
≥ 11.25 to < 20	≥ 135,000 to < 240,000		11.5 EER	$50	12.0 EER	$95
Large Commercial Air Cooled RTU and Split Systems						
≥ 20 to < 63	≥ 240,000 to < 760,000		10.5 EER	$30	10.8 EER	$60
≥ 63	≥ 760,000		N/A	N/A	10.2 EER	$60
Water and Evaporatively Cooled Air Conditioning Systems						
≥ 20	≥ 240,000		14.0 EER	$80	N/A	
Heat Pump Systems						
Air Cooled Heat Pump Systems						
< 5.4	< 65,000	Split	14.0 SEER & 12.0 EER & 8.5 HSPF	$70	15.0 SEER & 12.5 EER & 9.0 HSPF	$125
< 5.4	< 65,000	Packaged	14.0 SEER & 11.6 EER & 8.0 HSPF	$70	15.0 SEER & 12.0 EER & 8.5 HSPF	$125
≥ 5.4 to < 11.25	≥ 65,000 to < 135,000		11.5 EER & 3.4 COP	$50	N/A	
≥ 11.25 to < 20	≥ 135,000 to < 240,000		11.5 EER & 3.2 COP	$50		
≥ 20	≥ 240,000		10.5 EER & 3.2 COP	$30		

(EEAC 2015)

This measure's primary delivery channels are rebate programs, usually marketed through program administrator staff and heating, ventilating, and air conditioning (HVAC) contractor partners, or an upstream market program marketed through distributors. The programs provide:

- Rebates for units installed in commercial settings, typically paid on the basis of dollars-per-ton of cooling, which can vary by the efficiency level achieved.
- Rebates for residential units, which are often paid on a fixed rebate-per-unit basis to discourage oversizing, and to promote high-quality installation practices.

The rebates apply (1) at the time of normal replacement due to age or failure or (2) for new construction applications. Rebates are not usually offered for early replacements, except when unusually high use of air-conditioning occurs.

When a unit is installed in new construction or replaces an existing unit that has failed or is near the end of its life, the baseline efficiency standard it must meet is generally defined by local

[4] MassSave Cool Choice Program, offered in 2016-18 by all Massachusetts Program Administrators. http://ma-eeac.org/wordpress/wp-content/uploads/EEAC-CI-Innovation-Upstream-Memo-2015-12-30.pdf
Additional program examples:
https://www.premiumcooling.com/upload/2017_PECP_Equipment_Incentive_Schedule%20(FINAL).pdf;
https://focusonenergy.com/sites/default/files/HVAC_Plumbing_Catalog_v07_012017_web.pdf

energy codes, federal manufacturing standards, or ASHRAE Standard 90.1 for SEER-rated units (less than 5.4 tons) and IEER-rated units (5.4 tons or greater). This protocol assumes more efficient equipment of the same capacity runs close to the same number of hours as the baseline equipment. It does not cover:

- Early replacement retrofits
- Right-sizing initiatives
- Tune-ups
- Electronically commutated motor (ECM) retrofits on fans
- Savings resulting from installation of an economizer[5] or economizer controls, demand controlled ventilation, multi-unit controls, solar-power assistance, or energy recovery ventilation at the same time as installation of the new, high-efficiency equipment.

2.1 Programs with Enhanced Measures

Many program administrators offer other cooling measures in conjunction with higher EER/SEER/IEER cooling units. These measures include dual enthalpy economizers, demand controlled ventilation, and ECMs for ventilation fans as a retrofit or an upgrade option at the time of replacement.

Other programs, particularly residential, also focus on high-quality installation by requiring the work to meet Air Conditioning Contractors of America Quality Installation standards, which encompass proper duct sealing (ACCA 2015).

The evaluation methods addressed in this protocol do not include—on a standalone basis—savings resulting from these other measures. However, some overlap may occur with the evaluation, measurement, and verification (EM&V) of high-efficiency cooling system upgrades, particularly with demand controlled ventilation, ECMs, and dual enthalpy economizers.

2.1.1 Economizers

Economizers work by bringing in outside air for ventilation and cooling when outside conditions are sufficiently cool. In some jurisdictions, many of the new packaged or split systems have temperature or dry bulb-based economizers, as required by code or by standard practice. Evaluators can include units with temperature-based economizers in evaluation samples as a random occurrence as long as their occurrence in the sample is roughly the same proportion to their penetration in the population.

A dual-enthalpy economizer—a more sophisticated type, controlling both temperature and humidity—brings in outside air when the outside conditions are sufficiently cool and dry. These units tend to reduce the run hours of high-efficiency air conditioners as compared to units without economizers, thus reducing potential savings from more efficient units. Although dual-enthalpy economizers usually[6] are not required by code for these smaller sized units, some

[5] Most codes nationwide require basic economizers, such that baseline EFLH and measure EFLH should include free cooling, but measurement will likely reveal less than 100% functioning.
[6] Codes in California, Washington, and Oregon require advanced economizer controls in some applications.

utilities provide an incentive for them. More recently, some programs have provided incentives for adding more advanced digital economizer controls, which are similar to dual enthalpy economizers, but these controllers may provide additional compressor lockout and fan control. Units with advanced digital economizer controls also include automated fault detection diagnostics. If programs offer additional incentives for these more advanced economizers, savings for those measures should not be estimated using the protocol described here.

2.1.2 Demand Controlled Ventilation

Demand controlled ventilation (which uses a CO_2 sensor on return air to limit the intake of outside air to be cooled) can reduce the run hours for unitary and split systems. Units that receive rebates for demand controlled ventilation should not use this protocol, which assumes the equivalent full-load hour (EFLH) or load remains constant.

2.1.3 Right-Sizing

The savings estimated for this measure do not include the effects of right-sizing initiatives, which match outputs of cooling systems with cooling loads of facilities (thereby optimizing systems operations). The high-efficiency upgrade measure described here assumes both the base or code-compliant units and the high-efficiency units are the same size. Thus, the savings achieved through right-sizing initiatives must be determined using a more complex analysis method than is described here.

3 Savings Calculations

The calculation of gross annual energy savings for this measure, consistently defined by a number of technical reference manuals (TRMs) (MA Energy Efficiency Advisory Council Consultant Team 2015; United Illuminating Company and Connecticut Lighting and Power Company 2008; Vermont Energy Investment Corporation 2010), uses the following algorithms:

Cooling

Equation 1 (for units with a capacity of 5.4 tons or more)

$$\text{kWh Saved} = (\text{Size kBtu/hr}) * (1/EER_{baseline} - 1/EER_{installed}) * (EFLH_{cooling})$$

Equation 2 (for units with a capacity of fewer than 5.4 tons)

$$\text{kWh Saved} = (\text{Size kBtu/hr}) * (1/SEER_{baseline} - 1/SEER_{installed}) * (EFLH_{cooling})$$

Where:

Size kBtu/hr = cooling capacity of unit
$EER_{baseline}$ = energy efficiency ratio of the baseline unit, as defined by local code
$EER_{installed}$ = energy efficiency ratio of the specific high-efficiency unit
$SEER_{baseline}$ = seasonal energy efficiency ratio of the baseline unit, as defined by local code
$SEER_{installed}$ = seasonal energy efficiency ratio of the specific high-efficiency unit
$EFLH_{cooling}$ = equivalent full-load hours for cooling

While many efficiency providers currently use Equation 1 with EER for units of greater than 5.4 tons, the protocol recommends using the more accurate measure of seasonal efficiency, IEER, shown in Equation 3.

Equation 3 (for IEER)

$$\text{kWh Saved} = (\text{Size kBtu/hr}) * (1/IEER_{baseline} - 1/IEER_{installed}) * (EFLH_{cooling})$$

Where:

$IEER_{baseline}$ = integrated energy efficiency ratio of the baseline unit, defined to be minimally compliant with code, which is usually based on ASHRAE 90.1-2010
$IEER_{installed}$ = integrated energy efficiency ratio of the specific high-efficiency unit

Note that for many programs currently offered, only EER is required to qualify units 5.4 tons or greater. EER is not meant to represent annual efficiency and there is some error introduced by not using the IEER, but as of now there is no accepted general relationship between the two to use. It is recommended that all programs move toward using IEER or SEER for rebate qualification and energy savings estimates, and recording those values in the program tracking database. Peak demand savings are covered in Uniform Methods Project (UMP) Chapter 10;

however, in general we recommend using EER, which represents the system's full load efficiency, in any calculations of peak demand reduction.

For smaller units, SEER is almost always available, and it should be used for the calculation of annual energy savings. These formulas are consistent with ASHRAE Guideline 14–2014, although the guideline does not specify annual full load hours and focuses on the period of measurement.

Heating

Equation 4 (for units with a capacity of 5.4 tons or more)

$$\text{kWh Saved} = (\text{Size kBtu/hr}) / 3.413 \text{ kBtu/hr/kW} * (1/COPH_{baseline} - 1/COPH_{installed}) * (EFLH_{heating})$$

Equation 5 (for units with a capacity of fewer than 5.4 tons)

$$\text{kWh Saved} = (\text{Size kBtu/hr}) * (1/HSPF_{baseline} - 1/HSPF_{installed}) * (EFLH_{heating})$$

Where:

Size kBtu/hr = heating capacity of unit
$COPH_{baseline}$ = heating coefficient of performance of the baseline unit as defined by local code
$COPH_{installed}$ = heating coefficient of performance of the specific high-efficiency unit
$HSPF_{baseline}$ = heating seasonal performance factor of the baseline unit, as defined by local code
$HSPF_{installed}$ = heating seasonal performance factor of the specific high-efficiency unit
$EFLH_{heating}$ = equivalent full-load hours for heating

These formulas assume some simplifications: (1) baseline units and high-efficiency units are of equal size (that is, no downsizing or "rightsizing" due to increased efficiency); and (2) baseline and high-efficiency units have the same operating hours. Although this may not be the case for a given cooling or heating load, these simplifications have been determined reasonable in the context of other uncertainties.

4 Measurement and Verification Plan

When choosing an option, consider the following factors:

- The equation variables used to calculate savings
- The uncertainty in the claimed estimates of each parameter
- The cost, complexity, and uncertainty in measuring each of those variables.

When calculating savings for unitary HVAC, the goal is to take unit measurements as cost-effectively as possible, so as to reduce overall uncertainty in the savings estimate. Thus, use these primary components:

- Unit size
- Efficiency of the base unit and the installed unit
- Annual operating hours for energy savings
- Coincidence factor for demand savings.

4.1 IPMVP Option

The recommended approach entails two steps: (1) Use one of the equations provided above with manufacturer rated values for capacity and efficiency (using industry-approved methods); and (2) incorporate program-specific measured values for the operating hours. This approach most closely resembles International Performance Measurement and Verification Protocol (IPMVP) Option A: Partial Retrofit Isolation/Metered Equipment.

Option A can be considered the best approach for the following reasons:

- The key issue for replace-on-failure/new construction programs is the usage of baseline equipment, defined as the *current* code or prevailing standard. However, this cannot be measured or assessed for participating customers because, by definition, lower-efficiency baseline equipment was never installed. The unit replaced is often old and below current requirements and is not the appropriate baseline. A nonparticipant group installing baseline equipment could be used, but only one known study has attempted this to date (KEMA 2010). For most situations, finding valid nonparticipants through random-digit dialing and performing extensive metering is simply too costly, given the savings level this measure contributes to typical portfolios.[7]
- Regarding the use of pre/post-billing analysis (IPMVP Option C) for participants, the same issue applies—pre-installation does not represent the baseline. Even without using pre/post-billing analysis, one might try using monthly billing data to determine cooling energy for a facility and then calculate facility-level full-load hours for use in the equations. However, this method is not recommended because cooling electricity usage cannot be easily disaggregated from total monthly electric usage with the accuracy

[7] This generally represents a small percentage of total commercial and industrial portfolio savings; primarily due to code, most new equipment is already relatively efficient.

required. As more residential and small commercial customers get kilowatt interval data (hourly or smaller time intervals), estimating cooling hours from whole-building data may become more feasible for very simple cases, but such methods are error-prone; feasibility will depend strongly on building size and type, HVAC system configuration, and the profiles of other loads.

- Option D (Calibrated Simulation) in which savings are determined using building simulation, is also not a recommended evaluation approach for the measure. Option D involves developing an energy model to estimate energy use for a proposed building. Often the measures in this protocol replace individual units and developing a whole building model and calibrating can be too costly for each sample point. Option D is primarily intended for new construction projects, or major retrofits with multiple measures, where a whole building approach includes HVAC and other measures that affect the HVAC loads. The protocol uses Option A and is applicable to new construction if only evaluating the unitary HVAC measures.

4.1.1 Capacity

Measuring cooling capacity is extremely expensive and would only result in replicating information already provided in a manner overseen by a technical standards group (AHRI). Thus, for a unit's peak cooling capacity (size), use the manufacturer's ratings, as these have generally been determined through an industry-standard approved process at fixed operating conditions. Although some variation may occur in the output of individual rebated units, it is assumed that on average, units perform closely to AHRI ratings.

4.1.2 Efficiency Rating

For determining the efficiency levels of base units and installed units, an industry accepted standard alternative to *in situ* measurement is available through manufacturers' ratings. (Also, performing *in situ* measuring is extremely costly.)

4.1.3 Equivalent Full-Load Hours

The EFLH variable must be measured or estimated for the population of program participants. Operating hours are specific to building types and to system sizing and design practices. Typical design practice tends to result in oversizing (using a larger-than-needed unit). In general, the greater the oversizing, the fewer the operating hours, and the less efficiently a unit operates.

Two primary methods exist for developing hours of use for the equations in *Savings Calculations*—creating a building simulation or conducting metering. The recommended approach favors using some actual measurement rather than relying exclusively on simulation-based estimates.

Detailed building simulation prototype models can be developed for a wide variety of building types, system configurations, and applicable weather data. Such analysis usually results in an extensive set of look-up tables for operating hours listed by building type and weather zone. Various TRMs use this approach, including New York and California (TecMarket Works 2010; Itron, Inc. 2005). In California, DEER look-up tables contain 9,000 unique combinations of unit types, building vintages, climate zones, and building types.

This approach is used to establish deemed savings and program planning estimates, but it does not include measurements to account for oversizing practices or the types of building populations served by the actual programs. Thus, the recommended approach entails metering demand (kW) for a sample of units to develop EFLH estimates (KEMA 2010).

Note that the energy consumption of the compressor(s), condenser fan(s), and evaporator (i.e. supply) fan(s) are used to calculate the EFLH, but only when the compressor and condenser actually supply cooling or heating.

Measurement of energy consumption can be used to validate building simulation models. However, in practice, the cost of metering the sample sizes required for developing data for all building types and weather zones would be cost-prohibitive and thus has not been attempted. In a California study, results from approximately 50 units in three climate zones were used to develop realization rates to calibrate the simulation approach to metered data, but not to determine EFLH for combinations of building types, climate zones, and system types (Itron, Inc. and KEMA 2008).

Measuring energy consumption involves on-site inspections, where unit-level power metering is performed for a wide range of temperature, occupancy, and humidity conditions. The resulting data can be analyzed to determine energy consumption as a function of outdoor wet-bulb or dry-bulb temperatures. These data can be extrapolated to the entire year by using typical meteorological year (TMY) data.

Dividing annual energy consumption (kWh) by the peak rated kW serves as a proxy for EFLH. The peak rated kW is defined as a unit's peak cooling capacity at AHRI conditions in kBtu/hr and divided by the EER or the peak heating capacity at AHRI conditions divided by (COPH*3.412). Metering used to determine the annual kWh consumption should be based on either (1) a true power (kW) meter and integration of power over time; or (2) an energy meter, which performs the integration internally. Such metering should include the compressor(s), condenser fan(s), and supply fan(s). If true power kW or energy metering proves too costly, amperage data may be acceptable if they are supplemented with spot power measurements under a variety of loading conditions.

When taking measurements, consider these factors: (1) Use a random sample of units spread across building types and (2) stratify the sample by climate zone (if the territory has a wide range of temperature and humidity conditions) and unit sizes. Note that unit-size stratification may not be required if unit sizes fall within a narrow range. Please see UMP *Chapter 11: Sample Design Cross-Cutting Protocol* for additional details.

Although a sufficiently large random sample would likely capture the predominant building types of interest, we recommend checking distributions of building types in the sample relative to the population and then adjusting or redrawing the sample, as needed, if an adequate distribution does not result.

4.2 Verification Process

The key data to be verified are (1) the size of the unit rebated and (2) the nameplate efficiency of the installed unit. Verification can be performed through:

- A desk review of invoices and manufacturers' specification sheets (which should be required for rebate payment)
- An on-site audit of a sample of participants (usually the same participants selected for the end-use metering, discussed above)
- Request from program staff, distributors or contractors, and internet search to obtain manufacturer manuals and cut sheets with unit performance at varying conditions.

Cooling capacity and efficiency are measured by manufacturers under standard conditions; however, the EFLH is site-dependent and not measured. Thus, the major uncertainty arises in the EFLH, so metering should concentrate on that quantity.

If savings can be determined as a function of building types, then verification of building types on applications can be conducted through on-site visits or telephone surveys.

Baseline efficiency can be assumed to be that of a code-compliant unit in the service territory. Differences in efficiency between code-compliant units and standard practice would be reflected in the calculation of an appropriate net-to-gross ratio. [8]

4.3 Data Requirements

The minimum data required for evaluating a unitary HVAC rebate program are:

- Size (in Btu/hr or tons) of each unit installed
- Rated cooling efficiency (in EER, SEER, or IEER) of each unit installed
- Rated heating efficiency (in HSPF or COPH) of each installed unit, if applicable
- Assumed baseline efficiency for each category of units (from prevailing code or standard)
- Location of each unit, corresponding to specific weather station disaggregation used for analysis of metered data.

Metered data used in the evaluation consists of the EFLH developed for each weather zone, which is derived as the ratio of the annual kWh divided by the peak kW.

Using the appropriate equation in *Savings Calculations*, determine the savings for this measure with these data:

- The installed cooling capacity
- The EER, SEER, or IEER rating (from manufacturers' data) of the baseline unit and the installed unit
- The HSPF or COPH rating (from manufacturers' data) of the baseline and installed unit, if applicable

[8] Net-to-gross issues are addressed in UMP *Chapter 21: Estimating Net Savings – Common Practices*.

- The measured EFLH.

4.4 Data Collection Methods

Given the relative size of savings for this measure in a typical portfolio—one dominated by other higher-savings measures—the collection of data (which is comparatively costly) can best be conducted jointly with other program administrators in a state or region with similar weather conditions.

In the past 15 years, a number of studies have examined commercial unitary HVAC EFLH and load shapes of note (KEMA 2011; SAIC 1998; Itron, Inc. and KEMA 2008; KEMA 2010). Further, at least two studies have examined full-load hours of residential central air-conditioning systems (KEMA 2009; ADM 2009). The method this protocol recommends is based on work described in the Northeast Energy Efficiency Partnerships (NEEP) EM&V Forum study (KEMA 2011; Regional EM&V Methods and Savings Assumption Guidelines 2010), which, if conducted on a regional basis across multiple program administrators, balances rigor and cost.

As discussed, unit sizes and climate zones provide variables for developing a sampling framework. Large units tend to run for more hours and exhibit higher peak coincidence than small units (ranging from 3 tons to 15 tons). Large units also tend to use multiple compressors and are controlled differently than smaller, single-compressor units.

If a program predominantly rebates units smaller than 15 tons in size (or if the specific prescriptive program is limited to units smaller than 15 tons), only one size category is necessary. Similarly, if all units in the service territory or region studied have essentially the same temperature and humidity conditions (for example, one large city), sampling by climate zone is not needed.

Thus, if unit size and climate zone are not required sampling dimensions for representing the population, then sampling by predominant building type alone may be possible. Otherwise, sampling by combinations of climate zone, size, and building type may prove impractical.

4.4.1 Metering

Metering should capture integrated true root mean square kW power measurements at 15 minute intervals during at least half of the typical cooling season for the region, being sure to include either the spring or fall shoulder periods. If budget allows, metering should extend from the time units typically come on in spring until units are no longer needed in fall. Where budgets are constrained and timing allowed is not sufficient, the evaluator may meter for less time but should assure that the monitoring captures the preponderance of operating conditions to minimize the extent to which extrapolation must be performed outside the range of conditions captured. For high internal gain situations where cooling is needed year-round, metering should include some portion of the warmest weather and coldest weather months. If heating and cooling are to be derived for heat pumps, we recommend measurement of supply air temperature or control signal to indicate mode of operation. Full heat pump heating and cooling savings may require full year or near full year monitoring. This metering approach is consistent with the methods proposed in Normative Annex E of ASHRAE Guideline 14-2014.

Regardless of which metering intervals are used, data will be aggregated to one-hour averages for use in the model specified below, because publicly-available weather data are generally available in hourly formats.

This protocol is not designed to capture oversizing practices of the newly installed units, which requires more detailed cycling patterns (using 1-minute interval data) and goes beyond the determination of EFLH.

If budgets do not allow for measurement of kW using amperage and voltage measurements, using amperage measurements alone to determine EFLH and demand savings factors may be justified and is preferable over using values from studies conducted by other program administrators for similar climate zones and building types as described above. Direct kW measurements are preferable and the methods below assume kW measurements are taken. If amperage measurements are used, slight modifications to the formulas below for calculating EFLH are required.

The kW measurements should encompass the energy consumption of the compressor, condenser, evaporator, and supply fans. However, these measurements should only be used in the computation of the EFLH, when the compressor and condenser are actually running and supplying cooling (or heating, if applicable). The accuracy of kW measurements should be ± 2%, as recommended by Independent System Operator New England (ISO-New England, Inc. 2010).

After collecting the kW data, perform a unit-level regression of the unit power against predictor variables such as real-time weather data and whether the specific hour fell within the second or third hot day in a row. The predictor variables selected should provide the most significant independent variables for use as inputs to estimate the weather-normalized annual kWh consumption, and to extrapolate consumption outside the metering period. The result will be an 8760 kW load profile for that specific unit using the predictor variables. The following model functional form has been successfully used for this analysis in Northeast climates (KEMA 2011). Modifications to this model may be justified by the climate conditions and evaluation scope:[9]

$$L_{dh} = \alpha + \beta_{Ch} THI_{dh} + \beta_{w(d)} w(d) + \beta_{g(h)} g(h) + \beta_{2h} H_{2d} + \beta_{3h} H_{3d} + \varepsilon_{dh} \quad (2)$$

Where, for a particular HVAC unit:

L_{dh}	= load on day *d* hour *h*, day = 1 to 365, hour = 1 to 24 in kW
THI_{dh}	= temperature-humidity index on day *d* hour *h*
w(d)	= 0/1 dummy indicating day type of day *d*, Monday through Sunday and Holidays for eight dummy variables
g(h)	= 0/1 dummy indicating hour group for hour *h*, hour group = 1 to 24
H_{2d}	= 0/1 dummy indicating that hours in day *d* are the second hot day in a row

[9] For example, in hotter climates, the variable for consecutive hot days may not be needed or, in more humid climates, the dry bulb temperature and humidity may need to be separated

H_{3d} = 0/1 dummy indicating that hours in day *d* are the third or more hot day in a row

α, β_{Ch}, β_{Hh}, $\beta_{w(d)}$, $\beta_{g(h)}$ = coefficients determined by the regression

β_{2h}, β_{3h} = hot day adjustments, a matrix of coefficients assigned to binary variables (0/1) for hours defined for 2nd and 3rd consecutive hot days; matrix variables are unique to each hour in each hot day

ε_{dh} = residual error

The THI in °F can be defined as:

$$THI = 0.5 \times OSA_{db} + 0.3 \times DPT + 15$$

Where:

OSA_{db} = outside dry bulb temperature in °F

DPT = outside air dew point temperature in °F

Note that this particular functional form is just an example of what has been successfully used for commercial cooling. There is no preferred method specified in ASHRAE Guideline 14-2014. For heat pumps, we recommend the four-parameter change-point regression model or variable-base degree-day model, specified below. These models also assume hourly data.

Four-parameter change-point model:

Heating: $E = C - B_1(B_3 - T)^+ - B_2(T - B_3)^+$

Cooling: $E = C - B_1(B_3 - T)^+ + B_2(T - B_3)^+$

Where:

E = energy use

C = energy use at the change point

B_1 = coefficient or slope that describes the linear dependency on temperature below the change point

B_2 = coefficient or slope that describes the linear dependency on temperature above the change point

B_3 = change-point temperature

T = temperature for the period of interest

$^+$ = positive values only for the parenthetical expression, the lower bound is zero for the difference between T and B_3. This is not the same as the absolute value of the parenthetical expression.

Variable-base degree-day model:

$E = C + B_1\ (\mathrm{DD}_{BT})$

Where:

E	= energy use
C	= constant energy use below or above the change point
B_1	= coefficient or slope that describes the linear dependency on degree-days
DD_{BT}	= heating or cooling degree-days (or degree hours), which are based on the balance point temperature

However, this protocol is not suggesting that using this specific regression model is a requirement. Other examples of modifications include using a variable for the presence of economizers or using log functions with independent variables.

The success of the model should be measured by diagnostics such as signs for coefficients and comparison of measured power to modeled power via coefficient of variation - root mean squared error (CV-RMSE), R-square for the model, and the normalized mean bias error (NMBE)[10].

The following equation provides an EFLH calculation for the overall load shape (hourly load factor) or for each unit metered:

$$EFLH = \sum_{h=1}^{8760} \left(\frac{Estimated\ Hourly\ Load\ (kW)}{Connected\ Load\ (kW)} \right)$$

The connected load is defined as the unit's maximum kW recorded or peak cooling capacity at AHRI conditions in kBtu/hr divided by the EER. When performing the recommended analysis, the EFLH is defined for each range of temperatures for which there exists performance data including capacity and connected load. If the measurements are limited, a high and low range of capacity and efficiency is still recommended, as modern equipment for many capacities is two-stage or variable speed.

The HVAC unit's rated cooling capacity can be obtained from the unit make and model numbers, which should be required to be entered in the tracking system.

Although the EFLH is calculated with reference to a peak kW derived from EER, it is acceptable to use these EFLH with SEER or IEER. Some inconsistency occurs in using full-load hours with efficiency ratings measured at part loading, but errors in calculation are thought to be small relative to the expense and complexity of developing hours-of-use estimates precisely consistent with SEER and IEER.

[10] CV-RMSE and NMBE are fully defined with examples in ASHRAE Guideline 14. CV-RMSE measures deviations for each hour and thus measures model fit to the "load shape." NMBE measures the percent error over the entire performance period, in this case one year for annualized savings.

The EFLH for the population can be determined by multiplying the EFLH for each metered unit by the appropriate weights developed in the sample design (see Uniform Methods Chapter 11), reflecting that unit's contribution to the total population's cooling capacity.

Explicit 8760 load shape data are not always needed. This information, however, can be helpful for on-peak energy or demand savings calculations when (1) the time period in which the peak demand is being calculated differs among participants in a particular metering study or (2) the definition changes after primary data are collected. If the study has produced data for all hours of the year, these data can easily be reanalyzed for different on-peak energy and peak demand definitions.

4.5 Secondary Calculation

More extensive measurements than those described above may be justified when (1) typical operating conditions are significantly different than conditions for which the equipment has been rated or (2) the savings for this measure make up a significant portion of total portfolio savings. For example, extensive measurements may be appropriate in very hot and dry climates (such as the Southwest), where the dry-bulb temperature is often higher than the 95^{o}F used for EER ratings and the humidity is very low, compared to conditions for SEER ratings. Navigant (2010) has shown that performance in hot, dry climates differs significantly from manufacturers' standard conditions. DNV GL (2016 and 2017) performed IEER analysis using the HVAC Loadshape study to show potential different weighting that would lead to higher efficiencies for units in the Northeast with significant runtime at cool conditions where unit capacity and efficiency can be very high (Analysis not published).

Another complicating issue is performance at low loading for large units with multiple compressors running in parallel, or for units with variable-speed compressors. In such cases, low-loading performance is higher than expected from typical SEER ratings. If a part load rating is available that matches operating conditions reasonably well, use SEER or IEER in place of EER for simplified equations, calculating energy savings in conjunction with metered estimates of full-load hours.

In cases such as these, where more extensive measurement is justified, consider the following steps:

1. Meter equipment to determine runtimes in high and low stages of operation.
2. Aggregate and normalize runtime data for weather effects to create a typical hourly runtime shape that corresponds with a typical set of weather conditions.
3. Collect detailed performance data for a representative selection of equipment of various IEER/IPLV, EER, or SEER.
4. Calculate hourly kWh/ton using detailed performance data and runtimes for each hour for each piece of equipment.
5. Sum the hourly kWh/ton over the full year to calculate annual kWh/ton and then average hourly kWh/ton over the peak period to calculate peak kW/ton.
6. Fit a mathematical function to determine kWh/ton = f(SEER or IEER, EER) and kW/ton = f(SEER or IEER, EER).

7. Apply the mathematical functions for kWh/ton and kW/ton to the population's energy-efficient and baseline cases to determine savings for each piece of equipment.

An alternative for jurisdictions with detailed TRMs (such as New York) is the option used by Itron and KEMA in California, which involved measurement for a sample of units and development of a relationship between metered EFLH and that predicted by simulation models (Itron, Inc. and KEMA 2008). Expressed as a realization rate, such a relationship can be used for all unmetered sites to adjust simulation-based EFLH values. This alternative approach, however, is very expensive and, for equivalent funding, using the recommended approach can result in obtaining measurement data from five to 10 times more pieces of equipment. (Other measurement options are discussed in various ASHRAE publications [ASHRAE 2000; ASHRAE 2010; ASHRAE 2014].)

If all detailed measurements fall beyond an evaluation's available budget, program administrators can use available EFLH data from studies conducted for similar climate zones and building types. This approach, however, involves no actual measurements to reflect typical system sizing and design practices, building types, or weather in a region or service territory.[11]

[11] As discussed in the *Considering Resource Constraints* section of the "Introduction" chapter to this UMP report, small utilities (as defined under the U.S. Small Business Administration [SBA] regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

5 Sample Design

Evaluators will determine the required targets for confidence and precision levels, subject to specific regulatory or program administrator requirements and aligned with UMP *Chapter 11: Sample Design Cross-Cutting Protocol*. In most jurisdictions, the generally accepted confidence levels should be designed to estimate EFLH with a sampling precision of 10% at the 90% confidence interval. If attempting to organize the population into specific subgroups (such as building types or unit sizes), it may be appropriate to target 20% precision with a 90% confidence interval for individual subgroups, and 10% precision for the large total population.

In addition to sampling errors, errors in measurement and modeling can also occur. In general, these errors are lower than the sampling error; thus, sample sizes commonly are designed to meet sampling precision levels alone.

Sample sizes for achieving this precision level should be determined by estimating the coefficient of variation (CV), calculated as the standard deviation divided by the mean. Air conditioning and heat pump savings CVs generally range from 0.5 to 1.0,[12] and the more homogeneous the population, the lower the likely CV. After the study is completed, the CV should be recalculated to determine the actual sampling error of the metered sample.

As discussed, units should be sampled based on climate zones and unit sizes, if sufficient variation occurs in these quantities. Alternatively, the most prevalent building types can be sampled if the program administrator's database tracks building types accurately. One overall EFLH average can be developed if most units lie within a single climate zone and have a narrow range in capacity.

Many customers taking advantage of unitary HVAC rebate programs have multiple air-conditioning units rebated simultaneously. Consequently, the sampling plan must consider whether a sample can be designed for specific units, groups of units by size, or all units at a given site. It is also important to consider the resources needed to schedule and send metering technicians or engineers to a given site. Once those fixed costs have been incurred, metering multiple units at a site becomes an attractive option.

Decisions on how best to approach site (facility) sampling versus unit sampling depend on the degree of detail in the information available for each unit rebated. In many cases, rebate applications and tracking systems only record the total number of units in each size category, rather than the specific information on the location of each unit. For these instances, develop a specific rule that calls for random sampling of a fixed percentage of units at a given site.

Based on these considerations, sampling should be conducted per-customer site or application, with a specified minimum number of units sampled at a given site. A reasonable target is two or more units in each size category at each site with multiple units.

[12] At a CV of 0.5, the sample size to achieve a 10% precision with a 90% confidence interval is 67. At CV of 1.0, the sample size is 270. Program savings may vary less than EFLH when considering large geographic areas for multi-utility state, regional, or national studies.

6 Program Evaluation Elements

To assure the validity of data collected, establish procedures at the beginning of the study to address the following data issues:

- Procedures for filling in limited amounts of missing data
- Meter failure (the minimum amount of data from a site required for analysis)
- High and low data limits (based on meter sensitivity, malfunction, etc.)
- Units to be metered not operational during the site visit (For example, determine whether this should be brought to the owner's attention or whether the unit be metered as is.)
- Units to be metered malfunction during the mid-metering period and have (or have not) been repaired at the customer's instigation.

In addition to the raw data, the quality of an acceptable regression curve fit based on ASHRAE Guideline 14 (based on CV-RMSE, NMBE, R^2,) may further limit the sample. It is recommended to add to the sample an additional 10% of the number of sites or units to account for data attrition.[13]

At the beginning of each study, determine whether metering efforts should capture short-term measure persistence. That is, decide how the metering study should capture the impacts of non-operational rebated equipment (due to malfunction, cooling no longer needed, equipment never installed, etc.). For non-operational equipment, these could either be treated as equipment with zero operating hours, or a separate assessment could be done of the in-service rate.[14]

One key issue is how to extrapolate data beyond the measurement period for cooling-only units that may be left on after the primary cooling season ends. To address this and other unique operating characteristics, conduct site interviews with facility managers or homeowners (for residential units), as customers often know when units have been and are typically turned off for the season. These interview data can be used to omit non-typical data from the regression analysis indicating non-routine usage (e.g. cooling in the off-season), provided the customer can be certain the unit has not operated.

In analyzing year-round data from a mid-Atlantic utility, KEMA found that once the THI fell below 50^{o}F, most units shut off for the season. That information enabled KEMA to apply this rule to other sites in the NEEP EM&V Forum study, resulting in a more realistic estimate of fall and winter cooling hours than was obtained by applying only regression results (Regional EM&V Methods and Savings Assumption Guidelines 2010). If heating and cooling are to be derived, we recommend measurement of supply air temperature to indicate mode of operation.

[13] In KEMA's study for the NEEP EM&V Forum, approximately 9% of metered units were removed due to data validity problems (KEMA 2011).
[14] UMP *Chapter 6: Residential Lighting Evaluation Protocol* further discusses in-service rates.

7 Looking Forward

Since this protocol was first published in April 2013, there have been few, if any, metering studies on commercial unitary systems, and several studies for ductless mini-split and cold climate heat pumps for residential applications. Future evaluations are encouraged to include metering to provide valuable insight, since several TRM estimates may be based on aging estimates representing out-of-date equipment and milder climates. This protocol focuses on measuring total consumption and load shape, as these are typically the most uncertain parameters in HVAC measure savings. As of the date of publication of this protocol update (September 2017), there are lower cost options to meter HVAC loads at circuit breaker panels for both residential and nonresidential applications. Additional measurement options which are currently being studied for application to HVAC measures include:

- Efforts to further develop non-intrusive load monitoring options
- Efforts to use data from web-enabled "smart" thermostats
- Whole building hourly consumption analysis with advanced analytics.

In some situations, the efficiency change may require further scrutiny. These include using SEER and IEER in extreme climates, or early retirement where the existing unit efficiency is the relevant baseline for the remaining useful life of the removed equipment. This protocol does not include calculations for these situations. This protocol also does not address potential fan power savings during non-cooling and heating operation for space ventilation. Future protocols should:

- Develop coefficients to modify standard part load efficiency metrics to local climate and loads. Fairey et al. (2004) and new research is applicable to commercial buildings and IEER.
- Determine methods to estimate fan power savings for single zone variable air volume, two-speed and variable speed systems, and systems that reduce flow during ventilation-only operation (e.g., Advanced Digital Economizer Controls system).
- Consider protocol calculations for other measures that could be included in an HVAC measure in addition to the efficiency improvement, such as right sizing, adding economizers, and other load reduction or load shifting measures.

EM&V efforts can be used in larger studies to determine which measures perform best, or they can be deployed in targeted efforts. A larger effort would measure sufficient samples by technology, application, and climate, while a targeted effort would only sample the portions of a utility or state territory where participation is high or growing the fastest. There remains a challenge for TRMs with several full load hour combinations such as the simulation-based estimates in California and New York. This protocol recommends two options of either a large study designed for comparison across technology, application, and climate, or focusing on a specific combination and using the results of this protocol to produce end-use calibration targets. In this case, the most frequent combinations are calibrated and all other estimates are simulations of other combinations which should be proportionally correct, although not directly calibrated.

Although it may not be cost-effective to conduct a metering study solely for an HVAC rebate program, evaluators can leverage other in-home and commercial end-use metering efforts—such as those primarily designed to inform end-use disaggregation, which may initially be focused on residential applications. This approach may not yield a sufficient sample size for a primary metering study with one effort. However, if the data are collected in accordance with the guidance in Section 4.4, collective efforts among evaluators could yield robust samples that are sufficient to update regional estimates.

8 References

ADM. (November 2009). “Residential Central AC Regional Evaluation.” Prepared for NSTAR Electric and Gas Corporation, National Grid, Connecticut Light & Power, and United Illuminating. Air Conditioning Contractors of America (ACCA). (2015). *Standard 5 (ANSI/ACCA 5 QI-2015) HVAC Quality Installation Specification.*

Air-Conditioning, Heating and Refrigeration Institute (AHRI). (2008 with 2012 Addendum). *ANSI/AHRI 210/240-2008 with Addendum 2, Performance Rating of Unitary Air-Conditioning & Air-Source Heat Pump Equipment.* http://www.ahrinet.org/App_Content/ahri/files/standards%20pdfs/ANSI%20standards%20pdfs/ANSI.AHRI%20Standard%20210.240%20with%20Addenda%201%20and%202.pdf

AHRI. (2015). *Performance Rating of Commercial and Industrial Unitary Air-Conditioning and Heat Pump Equipment.* http://www.digeca.go.cr/sites/default/files/documentos/ahri_standard_340-360_2015.pdf

American Society of Heating Refrigeration and Air-Conditioning Engineers (ASHRAE). (2000). *Compilation of Diversity Factors and Schedules for Energy and Cooling Load Calculations,* ASHRAE Research Report 1093.

ASHRAE. (2014). *Guideline 14-2014 Measurement of Energy, Demand, and Water Savings*.

ASHRAE. (2010). *Performance Measurement Protocols for Commercial Buildings.*

California Public Utilities Commission. (2015-present). Database of Energy Efficient Resources (DEER) Updates. HVAC Measure updates in 2015, 2016, 2017, and 2018-19 Updates. http://deeresources.com/

Consortium for Energy Efficiency (CEE). (2016). *Commercial Unitary AC and HP Specifications, Unitary Air Conditioning Specification*. Effective January 12, 2016. https://library.cee1.org/system/files/library/7559/CEE_ComACHP_UnitarySpec2016.pdf

DNV GL. (2016 and 2017). 2013-14 and 2015 Upstream HVAC Impact Evaluation (HVAC1). Prepared for the California Public Utilities Commission. www.calmac.org

Fairey, P.; Parker, D.; Wilcox, B.; Lombardi, M. (2004). "Climate Impacts on Heating Seasonal Performance Factor (HSPF) and Seasonal Energy Efficiency Ratio (SEER) for Air Source Heat Pumps." ASHRAE Transactions, American Society of Heating, Refrigerating and Air Conditioning Engineers, Inc., Atlanta, GA, June 2004.

ISO-New England, Inc. (June 2010). *ISO New England Manual for Measurement and Verification of Demand Reduction Value from Demand Resources Manual (M-MVDR)*.

Itron, Inc. (December 2005). *2004-05 Database of Energy Efficient Resources (DEER) Update.* Prepared for Southern California Edison.

Itron, Inc. and KEMA. (December 31, 2008). *2004/2005 Statewide Express Efficiency and Upstream HVAC Program Impact Evaluation.* Prepared for the California Public Utility Commission, Pacific Gas & Electric Company, San Diego Gas & Electric Company, Southern California Edison, and Southern California Gas Company. www.calmac.org/publications/FINAL_ExpressEfficiency0405.pdf.

KEMA. (2009). *Pacific Gas & Electric SmartAC™ 2008 Residential Ex Post Load Impact Evaluation and Ex Ante Load Impact Estimates, Final Report.* Prepared for Pacific Gas and Electric. March.

KEMA. (February 10, 2010). *Evaluation Measurement and Verification of the California Public Utilities Commission HVAC High Impact Measures and Specialized Commercial Contract Group Programs 2006-2008 Program Year.* www.calmac.org/publications/Vol_1_HVAC_Spec_Comm_Report_02-10-10.pdf.

KEMA. (August 2, 2011). *C&I Unitary HVAC Load Shape Project*. Prepared for the Regional Evaluation, Measurement and Verification Forum facilitated by the Northeast Energy Efficiency Partnerships (NEEP).

MA Energy Efficiency Advisory Council (EEAC) Consultant Team. (December 30, 2015). *C&I Innovation – A Review of Upstream Offerings*. http://ma-eeac.org/wordpress/wp-content/uploads/EEAC-CI-Innovation-Upstream-Memo-2015-12-30.pdf

Navigant. (June 2010). "The Sun Devil in the Details: Lessons Learned from Residential HVAC Programs in the Desert Southwest." Presented at *Counting on Energy Programs: It's Why Evaluation Matters*. Paris, France: International Energy Program Evaluation Conference.

Premium Efficiency Cooling Program. (2017). *2017 Equipment Incentive Schedule.* https://www.premiumcooling.com/upload/2017_PECP_Equipment_Incentive_Schedule%20(FINAL).pdf

Regional EM&V Methods and Savings Assumption Guidelines. (May 2010.) Northeast Energy Efficiency Partnerships (NEEP) EM&V Forum.

SAIC. (1998). *New England Unitary HVAC Research Final Report*. Sponsored by New England Power Service Company, Boston Edison Company, Commonwealth Electric, EUA Service Company, and Northeast Utilities.

TecMarket Works. (October 15, 2010). *New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs- Residential, Multi-Family and Commercial/ Industrial Measures*. Prepared for the New York Public Service Commission. http://efile.mpsc.state.mi.us/efile/docs/16671/0026.pdf. A more recent 2017 version is also available at: http://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/72c23decff52920a85257f1100671bdd/$FILE/TRM%20Version%205%20-%20January%202018.pdf

United Illuminating Company and Connecticut Lighting and Power Company. (October 2008). *UI and CL&P Program Savings Documentation for 2009 Program Year*.

Vermont Energy Investment Corporation. (August 6, 2010). *State of Ohio Energy Efficiency Technical Reference Manual Including Predetermined Savings Values and Protocols for Determining Energy and Demand Savings*. Prepared for the Public Utilities Commission of Ohio.

Chapter 5: Residential Furnaces and Boilers Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

David Jacobson
Jacobson Energy Research
Providence, Rhode Island

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68561
September 2017

Contract No. DE-AC36-08GO28308

Chapter 5: Residential Furnaces and Boilers Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

David Jacobson
Jacobson Energy Research
Providence, Rhode Island

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68561
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter author wishes to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Ken Agnew and Jeremiah Robinson of DNV KEMA
- Arlis Reynolds and Matei Perussi of The Cadmus Group.

Suggested Citation

David Jacobson. (2017). *Chapter 5: Residential Furnaces and Boilers Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68560. http://www.nrel.gov/docs/fy17osti/68561.pdf

Acronyms

AFUE	annual fuel utilization efficiency
Btu	British thermal unit
CEE	Consortium for Energy Efficiency
DOE	U.S. Department of Energy
ECM	electronically commutated motors
EFLH	equivalent full-load hours
EM&V	evaluation, measurement, and verification
HVAC	heating, ventilating, and air conditioning
IPMVP	International Performance Measurement and Verification Protocol
M&V	measurement and verification
TRM	typical technical reference manual
UMP	Uniform Methods Project

Table of Contents

1 Measure Description

The high-efficiency boiler and furnace measure produces gas heating[1] savings resulting from installation of more energy-efficient heating equipment in a residence. Such equipment, which ranges in size from 60 kBtu/hr to 300 kBtu/hr, is installed primarily in single-family homes and multifamily buildings with individual heating systems for each dwelling unit. This protocol does not cover integrated heating and water heating units which can be used in lieu of space heating only equipment.

[1] High-efficiency equipment can also be fueled by propane; however, for this protocol to be applied, bills must be provided on a monthly basis.

2 Application Conditions of Protocol

Table 1 shows typical mid-level efficiency program rebate offerings for this measure.[2]

Table 1. Mid-Level Qualifying Efficiency and Rebate Values

Measure	Efficiency Requirement	Rebate Amount
Natural gas forced-air furnace	92% to 93.9% AFUE	$150
Natural gas forced-air furnace	94% to 95.9% AFUE	$300
Natural gas forced-air furnace	96% or higher AFUE	$400
Natural gas boiler	83.5% to 90.9% AFUE	$300
Condensing natural gas boiler	91% or higher AFUE	$500

A more aggressive program may offer the rebates shown in Table 2.[3]

Table 2. Higher-Level Qualifying Efficiency and Rebate Values

Measure	Efficiency Requirement	Rebate Amount
Natural gas forced-air furnace with ECM	96% or higher AFUE	$800
Natural gas forced-air furnace without ECM	95% or higher AFUE	$500
Natural gas hot water boiler	96% or higher AFUE	$1,500
Natural gas hot water boiler	90% to 95.9% AFUE	$1,000

The specific measure described in this protocol improves upon the efficiency of residential furnace and boilers in terms of the U.S. Department of Energy's (DOE's) annual fuel utilization efficiency (AFUE) rating. AFUE, the most widely used measure of seasonal thermal efficiency for residential-sized heating equipment, is defined as the amount of useful heat delivered from a unit into a heating system for distribution, compared to the amount of fuel supplied to the unit on an annual basis. Units with efficiency levels in excess of 90% generally rely on extracting additional energy—typically that lost up a flue—by condensing water vapor out of flue gas. Generally, this is accomplished using larger heat exchangers and a redesigned exhaust system to accommodate lower flue gas exit temperatures.

The measure primarily targets customers purchasing new equipment, usually for the following reasons:

- Acquiring a new home
- Converting to gas from oil or other fuel

[2] CenterPoint Energy's high-efficiency heating system rebate program offered in 2011. See www.centerpointenergy.com/services/naturalgas/residential/efficiencyrebatesandprograms/heatingsystemrebates/MN/.

[3] MassSave/GasNetworks 2012 High Efficiency Heating and Water Heating Rebates for Residential Customers. See www.masssave.com/~/media/Files/Residential/Applications%20and%20Rebate%20Forms/2012%20GN%20Rebate.ashx.

- Replacing equipment at the end of its normal life or upon failure
- Major remodeling of an existing home.

The program design assumes customers participating in a residential furnace and boiler program would purchase new equipment that meets applicable codes or standard practices. Therefore codes or standard practices provide the baseline from which savings can be calculated (rather than using the equipment being replaced as the baseline). The program seeks to encourage installation of higher-efficiency equipment by paying all or a significant portion of incremental costs for upgrading to such units.

Rebate programs, often used for such measures, are usually marketed through a utility (or other program administrator staff) and its heating and plumbing contractor partners. Typically, rebates are paid at a specified dollar amount per unit, depending on efficiency levels. (For residential equipment, size generally does not play a role, due to the narrow size range.)

Residential purchasers of new furnaces can also receive incentives for installing electronically commutated motors (ECMs) in place of standard efficiency motors on furnace fans or hot water distribution pumps. This protocol, however, does not cover ECMs, which primarily provide electricity savings. This protocol also does not cover add-on boiler control measures, such as outdoor temperature reset controls, as these often are used for retrofits of existing boilers.

Some comprehensive residential programs assist customers in determining the appropriate or "right" size of the unit to be installed relative to the predicted load of the home. As most residential boiler and furnace programs do not offer these services—and because the modeling becomes much more complex for programs that do—this protocol does not take into account the changes in capacity from such efforts.

3 Savings Calculations

Key issues in determining savings for this measure are:

- What data are collected at the time of installation or application for incentives?
- What data can be easily collected during an evaluation?
- What assumptions are made about baseline equipment-sizing practices?

As previously described, the installed unit's AFUE reflects its efficiency level. Typically, the AFUE rating is collected for each unit rebated, as incentive payments are contingent on receiving verification that the unit meets program requirements. However, the efficiency of the baseline unit is not typically tracked.

For determining unit-specific savings or overall average savings per unit, many common formulas calculating savings use unit size or "capacity" in their derivations. With air-conditioning units, the size or capacity ratings always are provided in cooling output (Btu/hr or tons) delivered from units. For heating equipment, however, both the rate of heat delivered from the system (that is, the output capacity) and the rate of energy the unit consumes (the input capacity) often are provided. Program administrators strive to be specific in their requests for the capacity ratings of incented heating units, however, the two ratings often are confused. Also, customers or plumbing and heating contractors sometimes fail to provide the information.

Input and output capacity ratings generally are provided as *peak* capacity and not annual average numbers represented by AFUE.

- For non-condensing boilers and for both condensing and non-condensing furnaces, the ratio of peak input and peak output come very close to the AFUE. Thus, nameplate data can approximate relative annual performance.
- For condensing boilers, peak capacity does not indicate annual performance well because units perform better at part-load conditions.

Thus, for the most efficient boilers (usually condensing units), it is not valid to assume the approximation of the ratio of rated peak input to output capacities is proportional to the AFUE. This difference carries implications regarding which formulas can be used to calculate savings.

Capacity values are needed for unit-specific calculations of gross savings, but when they are not supplied on rebate forms, program administrators often use the manufacturer-provided capacity information embedded in specific model numbers.[4] However, the capacity indicated in model number nomenclature usually provides the input capacity in kBtu/hr rather than the output capacity. Due to differences in the capacity information provided by program participants—and how this affects derivation of formulas for calculating savings—the recommended formula is presented in two forms; which one is used depends on the capacity value provided. (The

[4] For example, the York YP9C0**60**B12MP12C is 60,000 Btu/hr input capacity, and the York YP9C**100**C12MP12C is rated at 100,000 Btu/hr input capacity.

Appendix to this chapter provides derivations for calculating savings through these two methods.)

- The first derivation assumes data collected regarding unit size is input capacity.
- The second derivation assumes size data collected is output heating capacity.

Generally, input capacity (rather than output capacity) is more readily available, and the recommended formula for calculating savings is based on the following assumptions:

- Input capacity (Btu/hr) remains the same for the baseline unit and the installed unit.
- Annual full-load operating hours, operating hours, and output capacity differ for each unit. (This is a reasonable assumption, given that if input energy remains the same, and the installed unit more efficiently converts input energy to output energy, the more efficient unit will run for fewer hours.)

In these circumstances, use Equation 1 to calculate savings from a high-efficiency unit replacing a baseline-efficiency unit:

Equation 1

$$Savings_{b-e} = Capacity_{input\text{-}e} * EFLH_{e\text{-}installed} * [(AFUE_e / AFUE_b) - 1]$$

where:

$Capacity_{input\text{-}e}$ = peak heating input capacity of both the baseline and installed unit

$EFLH_{e\text{-}installed}$ = equivalent full-load hours of the installed high-efficiency unit

In some cases, program managers collect the output capacity (or what program managers interpret as output capacity).[5] The alternative formula for calculating savings has been based on an assumption that runtimes (and, therefore, output capacities) are the same for high-efficiency units and baseline units. However, input capacities of baseline units differ for base- and high-efficiency units.[6,7] That formula, based on the rated output capacity, is shown in Equation 2:

Equation 2

$$Savings_{b-e} = Capacity_{output} * EFLH * (1/ AFUE_b - 1/ AFUE_e)$$

where:

$Capacity_{output}$ = heating output capacity of both the baseline and installed high-efficiency unit

[5] On some rebate forms, the field simply says the “capacity”; it does not specify whether it is input or output capacity.
[6] This implies the same annual heating load on the home for the base and the installed unit.
[7] This assumes input capacities for the base and high-efficiency units are different (that is, the installer, knowing the unit is more efficient—or relying on the ratings—will install a unit with smaller input requirements for the higher-efficiency unit). This makes engineering sense but, again, it depends on whether input or output ratings are used.

EFLH = full-load equivalent hours of the baseline and installed high-efficiency unit

$AFUE_b$ = annual fuel utilization efficiency of the baseline code compliant/standard practice unit

$AFUE_e$ = annual fuel utilization efficiency of the high-efficiency unit

Note that the $Capacity_{output}$ * EFLH equals the annual heating (Btu or therms) loss of a home to be met by the furnace or boiler. It does not represent the peak design load (Btu/hr) used by HVAC contractors to size a system to meet the peak heating load.

An alternative formula for calculating savings uses results from multiplying Equation 2 by $AFUE_b/AFUE_b$ and noting $Capacity_{output} / AFUE_b = Capacity_{input\text{-}b}$.

Equation 3

$$Savings = Capacity_{input\text{-}b} * EFLH * [1 - (AFUE_b / AFUE_e)]$$

where:

$Capacity_{input\text{-}b}$ = heating input of the baseline unit

As the baseline unit's input heating capacity rarely is known, this equation is seldom used correctly. The equation is discussed here because it is sometimes used incorrectly, in that the output heating capacity is substituted for the base unit's input capacity. Given the issues discussed above regarding rated peak output capacity of condensing boilers not being related to the AFUE, do not use Equation 2 or Equation 3 when calculating the savings from condensing boilers.

4 Measurement and Verification Plan

When choosing an option, consider the following factors:

- The equation variables used to calculate savings
- The uncertainty in the claimed estimates of each parameter
- The cost, complexity, and uncertainty in measuring each of those variables.[8]

4.1 IPMVP Option

As gas energy efficiency programs have shorter histories than electric energy efficiency programs, considerably fewer impact evaluations have been conducted for either gas programs as a whole or for specific measures (such as replacements of boilers and furnaces). A thorough literature search for detailed evaluations of furnace replacement and boiler efficiency programs resulted in a very limited number of studies (NMR and Cadmus 2010) (KEMA 2009) (KEMA 2008). Thus, less information is available to inform the development of a recommended protocol, compared to many other measures.

Given the large sample sizes required and the high costs of gas submetering, it is not feasible to conduct direct gas submetering of a sufficiently large sample to represent varying types of equipment (boilers and furnaces with varying efficiency levels) and different home and homeowner characteristics. Fortunately, the possible end uses for gas in homes are limited, making disaggregation of whole-house gas billing data into heating and non-heating components very reliable. Consequently, the methods used to evaluate this program to date have involved whole-house gas billing data.

Option C is the recommended International Performance Measurement and Verification Protocol (IPMVP) option for this measure: whole-facility regression analysis combined with site-level data on the capacity and efficiency of an installed unit. The methods of Option C entail combining a billing analysis with the equations presented above, which produces the most useful results at a reasonable expense. The methodologies can provide updated deemed savings results or updated parameters for use in typical technical reference manual (TRM) equations, as listed in equations 1 through 3. This is based on:

- The potential variables in the equations used to calculate savings (as previously discussed)
- The cost and complexity in measuring each of those variables
- The availability and relevance of billing data.

The primary variables for determining savings for high-efficiency boiler and furnaces are:

1. The installed unit size or capacity in Btu/hr (either input or output)

[8] As discussed under the section *Considering Resource Constraints* of the "Introduction" chapter to this UMP report, small utilities (as defined under the U.S. Small Business Administration (SBA) regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

2. The AFUE rating of baseline unit
3. The AFUE rating of the installed unit
4. The annual equivalent full-load operating hours, determined from methods discussed below.

The key issue for evaluating time-of-replacement/replace-on-burnout/new construction programs is that baseline equipment cannot be measured or assessed for the same customer installing new equipment, as only high-efficiency units have been installed. Thus, the key challenges presented in evaluating this measure entail determining (1) what a customer would have installed in the program's absence and (2) how much energy the baseline equipment would have used.

The methods described below combine whole-building billing analysis with the savings equations provided above to calculate the evaluated gross savings for this measure.

4.2 Verification Process

The first step of the protocol entails verifying key program data collected on typical rebate forms, including the size (Btu/hr) and efficiency (AFUE) of the high-efficiency unit installed. Such data can be verified using a desk review of invoices and manufacturer specification sheets (which should be required for rebate payment) or through an on-site audit of a sample of participants to verify the quality of self-reported information. If efficiency and unit capacity are not collected for each participant, it is recommended that program application requirements be modified to include these important data.

Generally, the size and efficiency ratings for baseline units cannot be verified. However, the baseline efficiency is assumed to be the code-compliant AFUE rating in the service territory for a unit of the same size as the high-efficiency unit. Differences between the code-compliant units and the standard practice should be reflected in calculations of appropriate net-to-gross ratios. If the net-to-gross is not considered within the specific jurisdiction, use the efficiency noted in standard practice.

The standard installation practice for each category of furnaces and boilers can be determined through conducting detailed interviews with HVAC contractors and plumbers (when possible) and collecting shipment data from regional distributors.

4.3 Data Requirements

The key data to be collected for impact evaluations of furnace and boiler upgrade programs are:

- Type of unit (natural gas furnace, condensing hot water boiler, or steam boiler)
- Capacity of the unit in input or output Btu/hr, depending on the algorithm selected to calculate savings (as discussed, input capacity is preferred, and it is important to be explicit regarding whether the specified capacity is input or output)
- Efficiency of the installed unit in AFUE
- Assumed baseline efficiency for each type of equipment

- Type of housing unit (single-family, multifamily having one to four units, multifamily having more than four units)
- Location of each unit in terms of city or ZIP code and state, if multiple climate zones are analyzed (the location will be used to calculate heating degree days for weather normalization)
- Post-installation billing data for a minimum of 12 months (if available, a full 12 months of pre-installation data should be compiled for the preferred analysis method, discussed below).

4.4 Collecting Data

4.4.1 Capacity Ratings

For a unit's heating capacity, use ratings from the manufacturer's specifications, which generally are determined through Air-Conditioning, Heating, and Refrigeration Institute (AHRI)[9] and DOE-approved standards for input and output capacity. As information already has been provided in an industry-approved manner, measuring input or output capacities through metering would be redundant. Although some variation may occur in an individual rebated unit's capacity, it is reasonable to assume that, on average, a unit's performance will be close to the manufacturer's ratings.

As noted, an issue exists regarding the capacity (input or output) captured for each unit in the program tracking system and whether this can be easily determined during an evaluation. Because input capacity is more readily available—*and* for high-efficiency condensing boilers, the relationship between the two capacities does not equal AFUE—use Equation 1. (The basis for the methodology is discussed below.)

4.4.2 Efficiency Levels

Similar to capacity, the efficiency levels of baseline and installed units would be extremely costly and difficult to field-verify over the heating season. Use the information on labels and the AHRI ratings for efficiency (an industry-accepted standard available in an online directory).[10]

4.4.3 Equivalent Full-Load Hours of Operation

Most equations use the number of equivalent full-load hours of operation as a variable for calculating savings. Depending on the evaluation methodology selected (as discussed below), this variable is either calculated as a product of the billing analysis-based evaluation, or it is not used at all in determining average savings per installation (also described below).

In some evaluations, direct measurement of operating hours has been attempted by metering furnace fans, but the technique has not been widely used. As many furnaces and boilers currently have more than one stage, the fan and pump hours do not always indicate the full-load hours needed for a calculation using full capacity as a variable.

---

[9] Often listed as Gas Appliance Manufacturers Association (GAMA) in the manufacturer's literature. The Air-Conditioning and Refrigeration Institute and GAMA merged in 2007 to form AHRI.

[10] www.ahridirectory.org/ahridirectory/pages/home.aspx

5 Discussion of Methodology

The methodology used to calculate savings for each unit and, if required, to calculate the corresponding EFLH, begins with Equation 1, provided here again. This assumes that the input Btu/hr would be the same for the baseline unit and the installed unit and that annual full-load operating hours, EFLH, and output capacity could be different.

Equation 1

$$\text{Savings} = \text{Capacity}_{\text{input-e}} * \text{EFLH}_{\text{e-installed}} * [(\text{AFUE}_{\text{e}} / \text{AFUE}_{\text{b}}) - 1]$$

where:

$\text{Capacity}_{\text{input-e}}$ = heating input of both the baseline and installed unit in Btu/hr

$\text{EFLH}_{\text{e-installed}}$ = equivalent full-load hours of the installed high-efficiency unit

Assuming the gas used for heating = normalized annual heating consumption of the high efficiency (NAH_{e}), determined from a billing analysis (as discussed below), then:

Equation 4

- $\text{Savings} = \text{NAH}_{\text{e}} * [(\text{AFUE}_{\text{e}} / \text{AFUE}_{\text{b}}) - 1]$

Assuming the AFUE is both available for a high percentage of units installed *and* accurately represents the efficiencies of baseline units and installed units over the year, this formula, combined with sufficient post-installation billing data, allows calculation of savings using a billing analysis.

The analysis offers an advantage over a simple deemed savings formula with estimated capacity and AFUE, in that the billing analysis has been based on actual heating consumption data. Such consumption data reflect the home's size, the unit's capacity, the building shell's efficiency and the operational schedules.

The analysis must first develop post-installation, normalized annual heating consumption (NAH_{e}). Chapter 8: *Whole-Building Retrofit* protocol addresses the recommended approach for this process, discussing a two-staged approach based on individual premise analysis. That approach begins by developing premise-specific estimates of overall normalized annual consumption (NAC), which is the combination of the end-use consumption of heating and other gas-baseline load (such as cooking and water heating).

Step 1 (analyzing the individual premise) and Step 2 (applying the Stage 1 model) within Chapter 8: *Whole-Building Retrofit* protocol provide guidance on models and on how to derive overall NAC from model results. (See Equation 5.)

Equation 5

$$NAC_e = \alpha * 365 + \beta_H H_0$$

NAH_e provides the equation's heating-related component, shown in Equation 6.

Equation 6

$$NAH_e = \beta_H H_0$$

Where:

β_H = heating slope in therms or hundred cubic feet (CCF) of natural gas per heating degree day

H_0 = the average normal heating degree days

α. = non-heating usages in therms of CCF per day

Generally, premise-level NAH_e is aggregated to a program-average NAH_e for each category of boiler or furnace measure and then analyzed to develop an estimation of savings for each category.

Once NAH_e has been determined for each home or individual boiler or furnace studied, the savings can be easily calculated using Equation 4 *if* the AFUE is available for each installed unit *and* the assumed baseline AFUE is estimated.

Savings can be specified in a manner as granular as the participation data allow. For example, savings could be disaggregated into the following categories:

- Warm air furnaces with ECMs between 92% and 94% efficiency
- Hot water boilers between 88% and 92% efficiency and with input capacities between 60,000 and 80,000 Btu/hr
- Steam boilers more than 150,000 Btu/hr.

If an evaluation seeks to update variables in a TRM, use either Equation 1 or Equation 2:

Equation 1

$$Savings = Capacity_{input\text{-}e} * EFLH_e * [(AFUE_e / AFUE_b) - 1]$$

or

Equation 2

$$Savings = Capacity_{output\text{-}e} * EFLH_e * (1/AFUE_b - 1/ AFUE_e)$$

In each case, $EFLH_e$ can be determined by Equation 7:

Equation 7

$$EFLH_e = NAH_e / Capacity_{input\text{-}e}.$$

The equation used is determined by:

- Whether the program collects $Capacity_{input}$, $Capacity_{output}$, or both as part of the application and data collection process
- What kind of equipment has qualified for incentives.

As previously discussed, equations using output capacity do not work for condensing boilers due to relationships between rated output capacity and AFUE.

Because these equations do not work universally for all types of equipment *and* the input capacity often is embedded in the model number's nomenclature, Equation 1 is the preferred way to calculate average savings per unit, assuming AFUE estimates accurately capture relative differences in efficiency.

Steps for calculating savings for each category of furnace and boiler are:

1. Determine the annual post-installation heating consumption NAH_e
2. Multiply the NAH_e by the percentage of increase in efficiencies of installed versus baseline units.

If using a TRM of the form Equation 1 or Equation 3, determine the EFLH for that category of equipment and then use the equation with the capacity and installed efficiency of each unit installed to determine the saving of each unit. Alternatively, use the average capacity and average installed efficiency to determine the category average savings.

5.1 More Refined Approach

The approach presented above is limited in that it does not contain (1) an analysis of pre-versus-post changes in consumption resulting from a furnace or boiler replacement or (2) actual measurement of actual efficiencies. That is, the approach is not grounded in any measurement of change in consumption resulting from the purchase of a new unit; instead, it relies on the post-consumption data and the ratio of baseline to high-efficiency AFUEs.

The post-only billing analysis also does not capture any potential "take-back" effect. In this instance, take-back could occur when participants purchase a more energy-efficient model than the baseline unit that participants otherwise would have, and then they "take" some of the actual or perceived savings to increase their comfort through higher thermostat settings.

A simple pre/post analysis is not possible because pre-replacement consumption data do not supply the appropriate baseline for a time-of-replacement program. Pre/post analysis, however, results in the consumption change between the installed high-efficiency unit and the older existing unit.

If one can reasonably determine the efficiency of the replaced unit in terms of $AFUE_{replaced}$, savings can be estimated using the three AFUEs:

- $AFUE_{replaced}$ = AFUE of the unit that was replaced
- $AFUE_e$ = AFUE of the high-efficiency unit
- $AFUE_b$ = AFUE of the baseline efficiency unit.

The difference in normalized annual heating (NAH) between the existing or replaced unit and the high-efficiency unit ($\Delta NAH_{e\text{-}replaced}$) can be determined through a billing analysis of participants.[11] The *Pooled Fixed-Effects* approach section of Chapter 8: *Whole-Building Retrofit* protocol discusses the model specification producing an average $\Delta NAH_{e\text{-}replaced}$. As with the premise-level modeling, the model's heating-correlated parts capture heating consumption.

For the general pooled fixed-effects model, the key components are H_{im} (heating degree days), P_m (post-period indicator, capturing pre-post change) and I_{ki} (the measure indicator variable). These combine to estimate the change in heating consumption between pre- and post-installation periods. The change in normalized annual heating consumption is calculated as shown in Equation 8:

Equation 8

$$\Delta NAH_k = \gamma_{Hk} H_{0k} + \Sigma_q \gamma_{Hkq} H_{0k} x_{qk}$$

where the data, model structure, and estimation procedures are as described in Chapter 8: *Whole-Building Retrofit*.

The two-stage, site-level modeling approach discussed in Chapter 8: *Whole-Building Retrofit* can also provide a suitable estimate of average ΔNAH_k, which may be calculated as the difference in pre- and post- heating components of site-level models for participants and a comparison group. However, separate components of the site-level models are less stable than the overall NAC. Thus, for installations of furnaces and boilers without domestic hot water, ΔNAC_k should be close to ΔNAH_k and better determined than the heating-only ΔNAH_k.

Using equations for $\Delta NAH_{e\text{-}b}$ and for billing analysis-determined savings $\Delta NAH_{e\text{-}replaced}$, the following derivation provides an enhanced method for calculating savings, based on a change in consumption captured through billing analysis rather than through post-only consumption.

Assuming:

Equation 9

$$Savings_{e-replaced} = \Delta NAH_{e\text{-}replaced} = NAH_e * [(AFUE_e / AFUE_{replaced}) - 1]$$

[11] Again, the approach recommended for this process is discussed in the "Whole-Building Retrofit" protocol.

Equation 10

$$NAH_e = \Delta NAH_{e\text{-replaced}} / [(AFUE_e / AFUE_{replaced}) - 1]$$

Equation 4

$$Savings_{e-b} = NAH_e * [(AFUE_e / AFUE_b) - 1]$$

Equation 11

$$Savings_{e-b} = \Delta NAH_{e\text{-replaced}} * (AFUE_e / AFUE_b) - 1) / [(AFUE_e / AFUE_{replaced}) - 1]$$

Equation 12

$$Savings_{e-b} = \Delta NAH_{e\text{-replaced}} * (1/AFUE_b - 1/ AFUE_e) / (1/AFUE_{replaced} - 1/ AFUE_e)$$

The efficiency of the replaced unit, $AFUE_{replaced}$, can be determined through surveys of installation contractors. Ideally, the surveys would cover the age and efficiency of the measure. In many cases, contractors will not know the efficiency of the model replaced, however, the process of estimating the efficiencies can be helped by information regarding the age of the units, or examples of specific models, manufacturers, and capacities.[12]

The accuracy of this method is highly dependent on the quality of the $AFUE_{replaced}$ estimate. Contractors may tend to underestimate the efficiency of units replaced to justify the sale of more efficient units. This under estimate of the replaced unit efficiency would underestimate savings from going from a new baseline to high efficiency unit. When using a contractor survey, verify the responses with on-site visits in which the efficiency of the older unit being replaced can be assessed.

As with methods based exclusively on post-installation heating consumption, the savings for each unit can be determined using Equation 9, including only estimates of AFUE and post-installation billing data. Again, the average savings can be broken down as finely as participation data allow. Savings could be separated out for major equipment types (hot water boiler, steam boiler, and warm-air furnace, with or without ECM), efficiency (AFUE, condensing or non-condensing), and size categories (Btu/hr ranges), as listed in the typical program offerings above.

The preferred method does not lend itself to determining the EFLH to be used in typical TRM equations, but rather to calculating average therm or MMBtu savings by category. If use of a simplified TRM is necessary, Equation 13 or Equation 7 can be used with the average capacity of each unit to determine EFLH.

[12] Preston's guides provide a good resource for efficiency specifications on old units: www.prestonguide.com.

If average of the $Capacity_{input}$ is known for each category, then:

Equation 13

$$EFLH = Savings_{e-b} / [Capacity_{input} * (AFUE_e / AFUE_b) - 1)]$$

or

Equation 7

$$EFLH = NAH_e / Capacity_{input\text{-}e}.$$

6 Sample Design

In general, the evaluator will determine the required target confidence and precision levels, subject to specific regulatory or program administrator requirements. In most jurisdictions, the generally accepted level should be designed to estimate the category-level savings or EFLH at a precision level of 10% at the 90% confidence interval. That said, as no physical measurements are involved, this protocol seeks to use data from *all* participants receiving rebates. Consequently, sampling error will be as low as the availability of billing, capacity, and efficiency data permit. Traditional sampling will not occur, unless large data gaps emerge in efficiency or capacity. For the preferred method, using pre- and post-billing data and efficiencies only *and* assuming the installed AFUE is collected for each participant, the availability of billing data presents the only limitation.

The billing analysis itself will have errors in the development of heating consumption and changes in heating consumption, but the precision of those regression-based estimates can be calculated. The target for these estimates should be better than +/- 10% at a 90% confidence level. As the analysis generally includes *all* participants with available billing, the efficiency and capacity data-sampling errors are essentially eliminated, and the primary error results from the billing analysis and the assumptions in the development of the equations provided in this protocol.

Errors in the accuracy of efficiencies and capacities provided by manufacturers versus actual values in the field will not be determined as part of this protocol, given the costs of measurement, but they are assumed to be small, relative to errors in the billing analysis.

6.1 Program Evaluation Elements

At the study's onset, procedures need to be established for data validity. The key issues to address are:

- Clear determination whether capacity data collected are input or output
- The number of months of billing data from a site that are considered to be the minimum needed for analysis
- The procedures for filling in limited amounts of missing billing data.

6.2 Net-to Gross

A separate cross-cutting protocol to determine applicable net-to-gross is planned for Phase 2 of the Uniform Methods Project.

7 References

KEMA. (2008). *Puget Sound Energy's Residential Energy Efficient Furnace Program Impact Evaluation*.

KEMA. (June 11, 2009). *New Jersey's Clean Energy Program Residential HVAC Impact Evaluation and Protocol Review.* www.njcleanenergy.com/files/file/Library/HVAC%20Evaluation%20Report%20-%20Final%20June%2011%202009.pdf.

NMR and Cadmus. (2010). *High Efficiency Heating and Water Heating Equipment Process and Impact Evaluation*. Conducted for Gas Networks, a group of New England gas utilities offering energy efficiency programs.

8 Resources

Fels, M.F. (1986). "PRISM: An Introduction." *Energy and Buildings*. (9); pp. 5-18. www.princeton.edu/~marean/publications/prism_intro.pdf.

Massachusetts Program Administrators. (2011). *Massachusetts Technical Reference Manual for Estimating Savings from Energy Efficiency Measures 2012 Program Year—Plan Version*.

New Jersey Board of Public Utilities New Jersey Clean Energy Program. (July 2011). *Protocols to Measure Resource Savings*. www.njcleanenergy.com/files/file/Library/NJ%20Protocols%20Revisions%207-21-11_Clean.pdf.

Northeast Energy Efficiency Partnerships (NEEP) EM&V Forum. (2010). "Regional EM&V Methods and Savings Assumption Guidelines."

Vermont Energy Investment Corporation. (2010). *State of Ohio Energy Efficiency Technical Reference Manual*. http://amppartners.org/pdf/TRM_Appendix_E_2011.pdf.

9 Appendix

9.1 Simplified Formulas for Calculating Savings From Upgrading the Efficiency of a Residential Gas Furnace or Boiler

9.1.1 Constant Input Btu/hr for Baseline and Installed Units

The following applies when the input Btu/hr is available from the tracking database.

The major simplifying assumption is this: Input Btu/hr for the baseline and the high-efficiency unit would be the same, as is usually the case. Some contractors install smaller units if those units prove more efficient, but many installers use a unit with the same input Btu/hr size. For new construction, one must assume baseline and high-efficiency units would be the same size.

Assuming a building has the same annual heat loss Q_{loss}, regardless of the heating-unit efficiency, then:

Q_{loss} = annual heat loss in Btu

$Capacity_{input}$ = furnace or boiler input heat rate Btu/hr

Then:

$$Q_{loss} = Capacity_{input} * EFLH_{\mathbf{b}} * AFUE_{\mathbf{b}}$$

$$Q_{loss} = Capacity_{input} * EFLH_{\mathbf{e}} * AFUE_{\mathbf{e}}$$

Where:

$EFLH_{\mathbf{b}}$ = equivalent full-load run hours of baseline (hrs)

$EFLH_{\mathbf{e}}$ = equivalent full-load run hours of efficient unit (hrs)

$AFUE_{\mathbf{b}}$ = efficiency of baseline unit %

$AFUE_{\mathbf{e}}$ = efficiency of efficient unit %

Then:

$$EFLH_{\mathbf{b}} * AFUE_{\mathbf{b}} = EFLH_{\mathbf{e}} * AFUE_{\mathbf{e}}$$

$$EFLH_{\mathbf{b}} = EFLH_{\mathbf{e}} * AFUEe / AFUE_{\mathbf{b}}$$

Savings result from the difference in gas heating consumption between the baseline unit and the efficient unit:

$$\text{Savings} = \text{Capacity}_{input} * \text{EFLH}_{\mathbf{b}} - \text{Capacity}_{input} \times \text{EFLH}_{\mathbf{e}}$$

$$= \text{Capacity}_{input} * \text{EFLH}_{\mathbf{e}} * (\text{AFUE}_{\mathbf{e}} / \text{AFUE}_{b}) - \text{Capacity}_{input} \times \text{EFLH}_{\mathbf{e}}$$

$$= \text{Capacity}_{input} * \text{EFLHe} * ((\text{AFUE}_{\mathbf{e}} / \text{AFUE}_{b}) - 1)$$

To use the normalized annual heating (NAH) of gas for heating from billing data, via a degree day-based regression analysis or end-use metering, apply this equation:

$$\text{NAH}_{\mathbf{e}} = \text{Capacity}_{input} * \text{EFLH}_{\mathbf{e}}$$

Substituting the above into the savings equation produces:

$$\text{Savings} = \text{NAH}_{\mathbf{e}} * [(\text{AFUE}_{\mathbf{e}} / \text{AFUE}_{\mathbf{b}}) - 1]$$

So savings can be calculated using the above equation without the input heating capacity (if not known). Alternatively, the NAH_e can be divided by the input capacity to calculate an $EFLH_e$, which can be used with the efficiencies to calculate savings using:

$$\text{Savings} = \text{Capacity}_{input} * \text{EFLHe} * (\text{AFUE}_{\mathbf{e}} / \text{AFUE}_{b}) - 1)$$

9.1.2 Constant Output Btu/hr for Baseline and Installed Units

The following applies when output Btu/hr is available from the tracking database; however, it does not apply to condensing boilers.

AFUE = Useful Heat Delivered Out of Boiler or Furnace/ Gas Input Capacity = $\text{Capacity}_{output} / \text{Capacity}_{input}$

Savings = Change in input for a given heating load

Input Energy_b = Annual Heating Load / AFUE_b

Input Energy_e = Annual Heating Load / AFUE_e

Assuming annual heating loads served are the same for baseline and high-efficiency equipment, and *output* capacities ($\text{Capacity}_{output-e}$) of unit and hours are the same for each unit:

$$\text{Annual Heating Load} = \text{Capacity}_{output-e} * \text{EFLH}$$

Then:

$$\text{Savings} = \text{Input Energy}_b - \text{Input Energy}_e = \text{Annual Heat Load} / \text{AFUE}_b - \text{Annual Heat Load} / \text{AFUE}_e$$

$$= \text{Annual Heat Load} * (1/\ \text{AFUE}_b - 1/\ \text{AFUE}_e)$$

$$= \text{Capacity}_{output} * \text{EFLH} * (1/\ \text{AFUE}_b - 1/\ \text{AFUE}_e)$$

Rearranging:

$$\text{Savings} = \text{Capacity}_{output} * \text{EFLH} / \text{AFUE}_e * [(\text{AFUE}_{\mathbf{e}} / \text{AFUE}_{\mathbf{b}}) - 1]$$

Noting that:

$$\text{Capacity}_{output} / \text{AFUE}_e = \text{Capacity}_{input\text{-}e} \text{ and } \text{NAH}_{\mathbf{e}} = \text{Capacity}_{input\text{-}e} * \text{EFLH}_{\mathbf{e}}$$

Yields the same equations as above:

$$\text{Savings} = \text{NAH}_{\mathbf{e}} * [(\text{AFUE}_{\mathbf{e}} / \text{AFUE}_{\mathbf{b}}) - 1]$$

So again, savings can be calculated using the above equation, without requiring output or input heating capacity (if not known), or the NAC_e can be divided by the input capacity to calculate an EFLH, which can be used with the efficiencies and output capacity to calculate savings using:

$$\text{Savings} = \text{Capacity}_{output} * \text{EFLH}_e * (1/\ \text{AFUE}_b - 1/\ \text{AFUE}_e)$$

Chapter 6: Residential Lighting Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – December 2014

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Scott Dimetrosky, Katie Parkinson, and Noah Lieb
Apex Analytics, LLC
Boulder, Colorado

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68562
October 2017

Contract No. DE-AC36-08GO28308

Chapter 6: Residential Lighting Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – December 2014

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Scott Dimetrosky, Katie Parkinson, and Noah Lieb
Apex Analytics, LLC
Boulder, Colorado

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68562
October 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter authors wish to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Jennifer Conseco and Mimi Goldberg, DNV GL
- David Barclay and Kiersten von Trapp, NMR
- Dick Spellman, GDS Associates
- M. Sami Khawaja and Scott Walker, Cadmus.

Suggested Citation

Dimetrosky, S.; Parkinson, K.; Lieb, N. (2017). *Chapter 6: Residential Lighting Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68562. http://www.nrel.gov/docs/fy17osti/68562.pdf

Acronyms

CFL	compact fluorescent lamp
CV	coefficient of variation
DOE	U.S. Department of Energy
EISA	Energy Information and Security Act
EUL	effective useful life
GISL	general service incandescent lamp
GSL	general service lamp
HOU	hours of use
HVAC	heating, ventilating, and air conditioning
ISR	in-service rate
LED	light-emitting diode
NTG	net-to-gross
PCF	peak coincidence factor
TRM	technical reference manual
UMP	Uniform Methods Project
W	watt

Protocol Updates

The original version of this protocol was published in April 2013 and revised in January 2015. This chapter has been updated to incorporate the following revisions:

1. **Shift the focus to light-emitting diodes (LEDs)**. The previous (2015) chapter language and guidance focused almost exclusively on compact fluorescent lamps (CFLs). Due to the ubiquity of LEDs coupled with the removal of CFLs from both manufacturer and program administrator offerings, the current chapter has updated the guidance to reflect this transition to LEDs.
2. **Addressed Energy Information and Security Act changes**. Incorporated discussion of the U.S. Department of Energy (DOE) Final Rules on General Service Lamps (GSL) for the second phase of EISA, which includes expansion of GSLs, reflectors, and lifetime of lamps given post-2020 changes.
3. **Updated in-service rate (ISR)**. The focus of the ISR section shifted to reflect LEDs and was simplified for lifetime ISR.
4. **Hours of use**. Included new metering studies as examples.
5. **Value line LEDs.** This new section includes a discussion on whether to address value line LEDs as a baseline or net-to-gross issue.

Table of Contents

List of Figures

List of Tables

1 Measure Description

Despite increasing market and regulatory uncertainty, residential lighting continues to represent a significant share of ratepayer-funded energy efficiency electricity savings. Up until a few years ago, program administrators achieved most of these savings by promoting the purchase and installation of compact fluorescent lamps (CFLs), both standard “twist/spiral” bulbs and specialty CFLs such as reflectors, A-Lamps, globes, and dimmable bulbs. In the past several years, most energy efficiency programs have transitioned to promoting solid-state light-emitting diode (LED) lamps instead of CFLs. This transition will probably accelerate in 2017 and subsequent years because of the new ENERGY STAR® v2.0 lighting specifications that are effectively limited to LED lamps.[1]

The 2007 Energy Independence and Security Act (EISA) required that, from 2012 through 2014, the energy efficiency of most types of screw-base light bulbs improve by approximately 28%, as measured by the efficacy in units of lumens per watt (lm/W). EISA requirements took effect in phases, beginning with 100-watt equivalents in 2012, 75-watt equivalents in 2013, and 60-watt and 40-watt equivalents in 2014, eliminating the domestic manufacturing or importation of legacy incandescent lamps. The legislation also has a second phase (backstop) provision ensuring that the previous EISA requirements produce savings equal to or greater than an efficiency standard of 45 lm/W by January 1, 2020.

On January 18, 2017, the U.S. Department of Energy (DOE) issued the Final Rules on General Service Lamps (GSL) for the second phase of EISA (U.S. Department of Energy 2016).[2] These rules, in general, expand the definition of GSLs, extending the covered lumen range, base types, and shapes, as well as reducing the types of bulbs exempted. According to the rulings, these expanded bulbs will be subject to GSL efficiency standards, including the 2020 backstop, starting January 1, 2020. DOE did not, however, address whether the 45 lm/W will remain the 2020 standard or if a different standard will be applied.

Since EISA took effect in 2012, many lighting efficiency programs have continued to realize significant savings, but evaluating these programs has become increasingly complex since—as a phased-in legislation—EISA makes it difficult to determine the baseline as well as the measure lifetime (that is, whether or not savings will be realized after 2020).

Given new regulations, increased complexity in the market, and the general shift from CFLs to LEDs, this evaluation protocol was updated in 2017 to shift the focus of the protocols toward LEDs and away from CFLs and to resolve evaluation uncertainties affecting residential lighting incentive programs, including these:

[1] As of early 2017, the vast majority of the qualified lamps on the ENERGY STAR Lighting v2.0 product list were LEDs, and a number of manufacturers (for example, GE) had stated that they were exiting the CFL market. (ENERGY STAR 2016)

[2] According to the Department of Energy, General Service Lamps are defined as General Service Lamps (GSLs) include general service incandescent lamps (GSILs), CFLs, general service LED lamps, organic light-emitting diode lamps, and any other lamps that are used to satisfy lighting applications traditionally served by GSILs. GSLs are used in general lighting applications and account for the majority of installed lighting in the residential sector.

- Incorporation of the latest DOE rulings on GSLs, including impacts on baselines, exemptions, and measure lifetime
- In-service rates (ISRs)
- Cross-sector sales and leakage.

2 Application Conditions of Protocol

Program administrators typically deliver residential lighting measures through these four mechanisms:

- **Upstream buy-down/markdown.** The most common approach to achieving residential lighting savings is to offer "upstream" incentives to manufacturers (buy-down) or to retailers (markdown) that reduce the cost of CFLs and LEDs for consumers. Because this delivery mechanism offers the discount at the time of purchase (that is, at the point of sale), the customers are not required to complete an application or any paperwork.
- **Direct install.** Many program administrators who offer residential audit programs include the direct installation of CFLs or LEDs at the time of an audit. Most programs offer audits at either no cost or at a highly-discounted cost to the customer, and there is usually no additional cost for the installed bulbs.
- **Giveaways.** Several program administrators have provided CFLs or LEDs free of charge to residential customers through the mail, at customer service offices, or at events organized by community, religious organizations, or local government agencies. In some programs, the CFLs or LEDs are mailed to customers only upon request. In other programs, the CFLs or LEDs are distributed without prior customer request. The amount of customer information collected at the time of the giveaway events varies; some program administrators require full name and contact information and others require no information.
- **Coupons.** Some program administrators have relied on instant (point-of-sale) or mail-in coupons as the incentive mechanism for residential lighting products. These coupons typically require that customers provide their names and contact information to obtain the product at the discounted price or to receive the rebate.

Although this Residential Lighting Evaluation Protocol applies to all of these delivery mechanisms, the strategies for collecting and analyzing the data necessary to calculate the savings tend to vary. This protocol highlights and provides details about the strategies and approaches to data collection and analysis.[3]

Also, program administrators may need to prioritize their evaluation resources to determine combinations of measures and delivery strategies, based on criteria such as the contribution to savings and the assessed uncertainty of those savings estimates. (For example, uncertainty can occur with programs that have not been evaluated for a while or that have shifting baselines.)

[3] As discussed in *Considering Resource Constraints* in the Introduction chapter to this UMP report, small utilities (as defined under the U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

3 Savings Calculations

Evaluators can calculate gross first-year energy savings from residential lighting measures through a number of algorithms.[4] However, this protocol recommends the following general algorithms:[5]

$$kWh_{saved} = NUMMEAS * (\Delta W/1{,}000) * HRS * ISR * IE_e \quad (1)$$

$$kW_{saved} = NUMMEAS * (\Delta W/1{,}000) * PCF * ISR * IE_e \quad (2)$$

where:

kWh_{saved} = first-year electricity energy savings measured in kilowatt-hours

kW_{saved} = first-year electricity peak demand savings measured in kilowatts

NUMMEAS = number of measures sold or distributed through the program

ΔW = delta watts (baseline wattage minus efficient lighting product wattage)

HRS = annual operating hours

PCF = peak coincidence factor

ISR = in-service rate

IE_e = cooling and heating interactive effects

This chapter covers the recommended techniques for estimating each of these parameters, based on either primary or secondary data.

[4] As presented in the Introduction, the methods focus on energy savings and do not include other parameter assessments such as net-to-gross, peak coincidence factor (or demand savings), incremental cost, or measure life.
[5] Evaluators should use CFL and LED-specific input parameter values where primary or secondary data allow evaluators to distinguish between them.

4 Measurement and Verification Plan

Evaluators should calculate the savings from residential lighting measures through a mix of measured and estimated parameters. This protocol recommends this approach, which is similar to Option A of the International Performance Measurement and Verification Protocol (National Renewable Energy Laboratory 2002), because the values for some parameters (such as annual hours of use [HOU]) can be directly measured through metering. However, evaluators should estimate other parameters (such as delta watts for upstream lighting programs) through other techniques.

4.1 Number of Measures Sold or Distributed

The administrator (or a third-party implementation contractor) should track the number of measures sold or distributed through a program and compile this information in a database that contains as much detail as possible about the measures delivered. This information is helpful not only for verifying quantity but for calculating a number of savings parameters elsewhere in this protocol. For example, for each lamp sold or distributed through the program, tracking data should include these:

- Product shipment dates from manufacturer to retailer, where applicable
- Detailed product information such as:
    - Bulb type (CFL, LED)
    - Wattage (three-way bulbs should include all wattage values)
    - Style and features (twist/spiral, reflector, A-Lamp, globe, dimmable, base type)
    - Manufacturer, model number, and product identifier (universal product code or stock keeping unit code)
    - Rated lumens
    - Rated life hours
    - Equivalent incandescent wattage, if available
    - Date of retail sale, if available
    - ENERGY STAR qualification
- Number of products incented (number of packs and number of bulbs per pack)
- Date incentive paid
- Dollar value of incentives paid
- Location where products were sold (including retailer name, address, city, state, and ZIP code)
- Final retail sales price of product, if available
- Company contact information (store manager or corporate contact name and phone number).

For programs using other delivery strategies, administrators should collect similar details. For example:

- An audit program would typically require the numbers and types of products installed, the wattage of the replaced bulb and location (room type), the date of installation, and customer contact information.
- A giveaway program would typically require at least the customer contact information, the quantity and type of product given away, and the detailed product information previously listed.

At a minimum, the evaluation should include a basic verification of savings, whereby the evaluator first sums the detailed transactions then attempts to replicate the calculation of total claimed savings for the specific period, such as a program year or cycle, during which the savings were claimed.

Evaluators should treat discrepancies between the claimed and verified number of measures as adjustments to the number of program measures. In other words, if the number of measures claimed by a program administrator does not match the detailed tracking data, the evaluator should first attempt to resolve the discrepancy with the administrator (perhaps the evaluator received incomplete records) and, if unable to resolve, should regard the amount recorded in the tracking data as the correct number.

4.2 Delta Watts

The difference between the wattage of the efficient lighting measure and the wattage of the assumed baseline measure is the delta watts. As noted, administrators should enter the wattage of the efficient measure in the program tracking database regardless of the program delivery mechanism.

Where possible—such as with direct install programs—the implementation contractor should record the wattage of the particular lamp that the program measure replaces.[6] Typically, this is done at the time of the audit, when auditors replace the existing measure with the efficient measure. However, this is not possible for most program delivery strategies, so evaluators often need to estimate baseline wattage. The baseline assumptions need to incorporate the transition to EISA standards that began in 2012 and further revised based on the DOE Final Rules on GSLs, issued January 18, 2017.

4.2.1 Approaches for Estimating Baseline Wattage

Recent studies have used these approaches for estimating baseline wattage:

- **Self-report.** Evaluators use customer surveys conducted after the installation to collect information about the wattage that consumers used before installation of the energy-efficient lighting.

---

[6] The baseline lamp typically has a much shorter lifetime than the retrofit lamp and the baseline may shift over the life of the retrofit lamp (particularly because of EISA).

- **In-home inspections to examine wattage of lamps in equivalent fixtures.** The implementation contractor examines the labeled wattage of bulbs in similar fixtures in each home to estimate the wattage the consumer used before the energy-efficient lighting was installed.
- **Multipliers.** Evaluators assume the baseline to be a multiple—for example, three or four times the wattage—of the efficient measure; thus, the evaluator will use one value (a single multiplier) across all program bulbs.
- **EISA lumen equivalence.**[7] EISA standards require that lumen ranges and assumptions about the equivalent wattage of incandescent lamps be specified on all retail lamp packaging (see Figure 1). Evaluators use the EISA-based lumen equivalency tables to determine the baseline wattage (examples are provided in Section 4.2.2).
- **ENERGY STAR lumen equivalence (manufacturer rating).**[8] Most energy-efficient lighting products prominently list replacement wattage assumptions on the box (Figure 1), and ENERGY STAR guidelines require these bulbs to use specific baseline wattages based on lumen bins.[9] The Energy Labeling Rule requires manufacturers to include detailed information about lamp brightness (lumens) and efficacy as part of the "Lighting Facts" label.[10] Evaluators use the actual equivalent rated wattage on the packaging.

Figure 1. Example of manufacturer-rated baseline wattage

---

[7] EISA bins are provided in the legislation online at Lightopedia. http://www.lightopedia.com/_files/eisa/energy-independence-and-security-act-of-2007.pdf.
[8] ENERGY STAR bins are provided on page 13 of the ENERGY STAR Lamp Specifications. https://www.energystar.gov/sites/default/files/ENERGY%20STAR%20Lamps%20V2_0%20Revised%20OCT-2016_1.pdf, page 13
[9] ENERGY STAR Lamps V1.0 requires a standard manufacturer baseline rating scale based on brightness (lumens) and bulb shape. Detailed specifications are available online. ENERGY STAR. "Certified Products." Available online: https://www.energystar.gov/products/specs/lamps_specification_version_1_0_pd
[10] Information about this rule is available online at: http://www.energystar.gov/index.cfm?c=cfls.pr_cfls_lumens. (Federal Trade Commission 2013).

Table 1 lists the strengths and limitations of each of these approaches.

Table 1. Strengths and Limitations of Alternative Delta Watts Estimation Approaches

Approach for Estimating Baseline Wattage	Strengths	Limitations
Customer self-report	Captures customer intentions and bin shifting*	Potentially low recall and social desirability bias
In-home inspections examining equivalent fixtures	Actual recording of baseline wattage for existing measures	Difficult to identify equivalent fixtures and high cost to conduct statistically representative on-site study. In addition, the existing in-home stock of lighting may not represent the actual delta watts that are available through retail purchases
Multipliers	Low effort and low cost; accuracy derived from empirical program data and, perhaps, better funded studies	Determining the appropriate multiplier for the program is difficult without basing it on another approach, or relying on other studies. The resulting estimate can be biased depending on the distribution of bulb type and wattages
EISA lumen equivalence	Widely available and relatively inexpensive to implement. In some cases, matches the marketed baseline wattage or matches up with EISA standards	May provide conservative estimate in cases where marketed baseline wattage exceeds rated lumen output
ENERGY STAR lumen equivalence (manufacturer-rated baseline wattage)	Widely available and relatively inexpensive to implement. Data based off wattage rating on package, which is often prominently displayed on the product. Approach is consistent with ENERGY STAR v2.0 specification	May not match the EISA lumen bins or be adjusted for EISA (that is, uses legacy bulb wattages)

*Bin shifting occurs when consumers do not replace bulbs with the same comparable wattage as the previous bulb (see Section 4.4).

The lumen equivalency bins for EISA legislation do not align with the ENERGY STAR lumen bins, further complicating the assessment of baseline wattages. This inconsistency results in EISA baselines varying from those noted on bulb packaging (see Figure 2). The recommended approach with how to deal with this inconsistency is reviewed in Section 4.4.

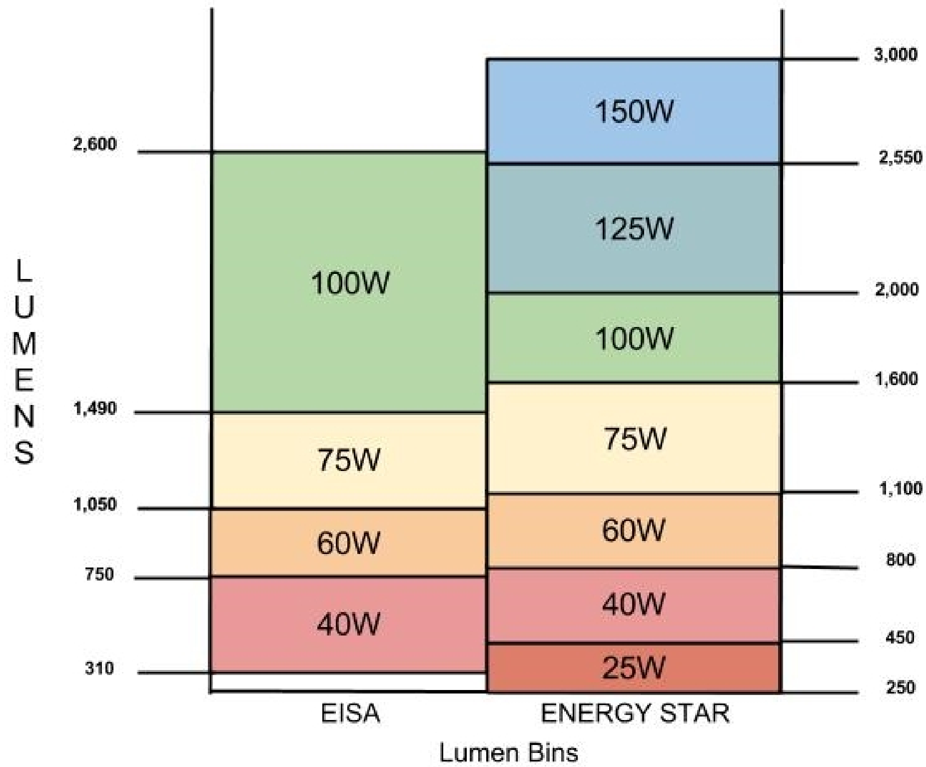


Figure 2. Baseline wattage by lumens, EISA versus ENERGY STAR

4.2.2 Recommended Approach

Consumers are more likely to purchase bulbs based on the rated baseline equivalent wattage rather than on the lumens.[11] Thus, for direct-install programs, the implementation contractor should collect baseline wattage information when the measure is installed. Where baseline information cannot be collected, the Residential Lighting Evaluation Protocol recommends using an adjusted ENERGY STAR lumen equivalency rating (manufacturer-rated baseline wattage) and then adjusting these estimates for the EISA requirements. The protocol recommends this approach because the manufacturer-rated baseline wattage for an ENERGY STAR bulb must be based on ENERGY STAR lumen categories. This approach incorporates EISA requirements, which are based on lumen output.

For studies that have sufficient budget to screen for a statistical sample of recent CFL or LED purchasers, evaluators may use the self-report approach to estimate delta watts (as well as other purchase attributes including location and price). This protocol recommends, however, that the customer self-report approach apply these time limits (from the time the consumer purchased the bulb):

[11] Recent studies have shown that consumers are still largely unaware of lumens. For example, a study from New York (NMR Group, Inc. 2014, NYSERDA) found that only 57% of respondents had even heard of the term *lumens* and, of those, more than 80% could not say how many lumens a 60-W bulb uses. A 2015 NMR study (NMR Group, Inc., 2015) from Connecticut demonstrated that only 54% of respondents had heard of lumens.

- A maximum three-month window for standard spiral CFLs or standard A-lamp LEDs
- Up to six months for specialty CFLs and LEDs, as these have far lower purchase incidence but represent larger purchase decisions.

When consumers do not replace bulbs with the same comparable wattage as the previous bulbs, this is called *bin shifting*. For example, a consumer may replace a 60-watt bulb with a 75-watt equivalent. Consumers can bin shift to higher- or lower-than-expected wattages. The self-report sampling approach offers the advantage of capturing consumer bin shifting, although there is little evidence that consumers bin shift when purchasing efficient lamps (Navigant et al. 2012).[12]

EISA legislation, as originally drafted, did not apply to all bulb types, which required evaluators to establish whether a bulb is exempt from EISA requirements. Therefore, to calculate savings prior to 2020, evaluators need to classify bulbs by shape, base type, lumens, and specialty features. Commonly used pre-2020 EISA-exempt bulbs include:

- Three-way bulbs
- Globes with ≥ 5-in. diameter or ≤ 749 lumens
- Candelabra base bulbs with ≤ 1049 lumens.[13,14]

The baselines for exempt bulbs should match the manufacturer-rated wattage (Column C in Table 2 and Table 3).

When synchronizing evaluated baselines to those noted on bulb packaging, it is important to be aware that the recommended lumen equivalencies differ for standard and specialty bulb shapes, in line with ENERGY STAR labeling requirements. Table 2 provides the assumed baseline wattage based on lumen range for GSL lamps (medium screw-base bulbs that are not globe, bullet, candle, flood, reflector, or decorative shaped). Evaluators can use the manufacturer-recommended baseline wattage for bulbs with lumens outside the lumen values shown in the table. Baselines in Table 2 apply to twist/spiral lamps and A-Lamps, and incorporate EISA phase-in periods through 2014. The baseline wattages listed in these tables reflect first-year savings, as well as savings up through 2020. The protocol recommendations for handling post-2020 savings are discussed in greater detail below.

[12] Navigant et al. (2012) found that only 2.6% of purchased CFLs might have been a different equivalent wattage than the incandescent bulbs they replaced.
[13] See EISA legislation for the full list of exemptions.
[14] Flood and reflector lamps have separate EISA requirements that took effect in July 2012. The flood- and reflector-specific lm/W requirements should be used as the baseline for any program equivalent lamps.

Table 2. GSL Estimated Baseline Wattage for Lumen Equivalencies

Minimum Lumens (a)	Maximum Lumens (b)	Incandescent Equivalent Wattage	
		Baseline (Exempt Bulbs) (c)	Baseline (Post-EISA) (d)
2,000	2,600	150	72
1,600	1,999	100	72
1,100	1,599	75	53
800	1,099	60	43
450	799	40	29
310	449	25	25

Table 3 provides the assumed baseline wattage—based on lumen range—for specialty and decorative-shaped lamps. Evaluators can use manufacturer-recommended baseline wattage for bulbs with lumens outside the values shown in Table 3. Specialty lamps are medium screw-base bulbs that are globe, bullet, candle, or decorative shaped.[15] Baselines in Table 3 incorporate EISA requirements.

Table 3. Specialty Lamp Estimated Baseline Wattage for Lumen Equivalencies

Lumen Bins		Incandescent Equivalent Wattage	
Decorative Shape (a)	Globe Shape (b)	Baseline (Exempt Bulbs) (c)	Baseline (Post-EISA) (d)
	1,100–1,300	150	72
	650–1,099	100	72
	575–649	75	53
500–699	500-574	60	43
300–499	350–499	40	29
150–299	250–349	25	25
90–149		15	15
70–89		10	10

Table 4 provides the EISA baseline calculations required for directional (reflector) lamps. Directional lamps include BR, ER, and BPAR lamps; reflector lamps between 2.25 inches (R18) and 2.75 inches (R22) in diameter; and lamps that have a rated wattage of 40 watts or higher. Directional lamps that are currently exempt from the requirements include BR30, BR40, and ER40 lamps rated at 65 watts (65BR30 and 65BR40 are exempt); ER30, BR30, BR40, and ER40 lamps rated at 50 watts or less; and R20 lamps rated at 45 watts or less. Although the data in the table can be used as a general reference for estimating reflector baseline equivalent wattages, this protocol acknowledges that the bulb characteristics required for this calculation are often unavailable, and even when available, produce minimum lumens per watt that are not realistically available for consumers. Some technical reference manuals (TRM) provide specific guidelines on the lumen bin equivalent wattages and evaluators may use the associated TRM equivalency tables for reflector baseline calculations as a proxy (Illinois Energy Efficiency Stakeholder Advisory Group 2017, Arkansas Public Service Commission 2016), as this is a more simplified approach. Evaluators should be wary of relying on the TRM reflector tables as the

[15] Bulb shapes that fit into this category are B, BA, C, CA, DC, F, and G lamp shapes.

most critical factor for reflector baselines are the general on-the-shelf availability of the lumen-equivalent bulbs.

Table 4. Directional (Reflector) Lamp Estimated Calculation for Baseline Wattage and Lumen Equivalencies

Lamp Wattage	Lamp Type	Diameter	Voltage	Calculation	Minimum Lumens per Watt
40W-205W	Standard spectrum	> 2.5" (PAR30, PAR38, BR30, BR40, ER30, ER40)	≥125 (130V)	6.8 x (Lamp watts ^0.27)	18.4 - 28.6
			<125 (120V)	5.9 x (Lamp watts ^0.27)	16.0 - 24.8
40W-205W	Modified spectrum	> 2.25" & 2.5" R20 & PAR20)	≥125 (130V)	5.7 x (Lamp watts ^0.27)	15.4 - 24.0
			<125 (130V)	5.0 x (Lamp watts ^0.27)	13.5 - 21.0

Evaluators should calculate baseline wattage for each lamp in the tracking database. Therefore, an evaluator should calibrate the total estimated delta watts to the actual type and number of measures sold or distributed through the program.

4.3 Calculating Lifetime Savings Post-2020

4.3.1 Changes to the EISA Post-2020 Legislation

The DOE Final Rules on GSLs, issued January 18, 2017, include two primary sets of rules: one that focused solely on reflector bulbs, and one focused more generally on GSLs (U.S. Department of Energy 2016). These rulings serve to update the definition of GSLs and assess the types of bulbs exempted in the current efficiency legislation. At a high level, these rulings expand the definition of GSLs, extending the covered lumen range, base types, and shapes, as well as reduce the types of bulbs exempted. According to the rulings, these expanded bulbs will be subject to GSL efficiency standards, including the 2020 backstop, starting January 1, 2020. Specifically, the ruling includes the following:

- **Reflector exemptions:** Reflector bulbs will no longer be exempt. The following three reflector lamp types (which represent the vast majority of reflectors) are no longer exempt from GSL standards: lamps rated at 50 watts or less that are ER30, BR30, BR40, or ER40 lamps; lamps rated at 65 watts that are BR30, BR40, or ER40 lamps; or R20 incandescent reflector lamps rated 45 watts or less.
- **Lumen maximums:** The lumen maximum subject to the EISA GSL definition has been expanded to 3,300 lumens (previously 2,600).
- **Base-type exemptions:** All standard bulb bases will be included (small screw-base and candelabra).
- **Other exemptions:** Three-way, decorative (including globes <5 inch, flame shapes, and candelabra shape), T-lamps (≤40w OR ≥ 10 inch), vibration service, rough service, and shatter-resistant bulb exemptions are to be discontinued. These bulbs will be subject to GSL efficiency regulations starting January 1, 2020.

These rules do not impose or amend efficacy standards for general service lamps; they are not addressing the 45 lumen/watt backstop requirement at this time, but maintain the option to do so later. The new ruling also made clear that a sell-through period may be expected, stating that: "it shall not be unlawful for a manufacturer to sell a lamp which is in compliance with the law at the time such lamp was manufactured. DOE expects it would interpret and apply the backstop with [this]… in mind" (U.S. Department of Energy 2016).

4.3.2 Calculating Post-2020 Savings

Bulbs expected to be in use in 2020 and beyond will be affected by the EISA backstop provision mentioned in Section 1. The life-cycle savings of efficient lamps, therefore, needs to account for a dual baseline:

- **Period 1:** Savings prior to the EISA backstop provision, which are based on the assumptions outlined above
- **Period 2:** Savings after the backstop provision, which are currently based on the 45 lm/W efficacy standard, and include many of the previously exempt lamps.

Although there are a few ways to account for this baseline shift, the Uniform Methods Project (UMP) protocols recommend applying a "sunset" year where savings can be claimed, to be determined by the period in which consumers are unlikely to find an alternative other than LED lamps. This sunset year could exceed 2020 for a few reasons, including:

- **Sell-through:** Although the original EISA provision had a hard stop on sales on January 1, 2020, as noted above, the latest rulemaking implies a sell-through period will be allowed.
- **Enforcement:** The federal government prohibited any funds being used for the EISA 2012–2014 phase-in enforcement, and similarly it is unlikely the 2020 provision will have enforcement.
- **Political uncertainty:** The entire EISA 2020 backstop could be overturned depending on which parties control the executive and legislative branches of government in the future.
- **Halogen burn-out period**: Even if the January 1, 2020 EISA backstop remains in place with enforcement, in theory an EISA compliant halogen could still be purchased in 2019, and that halogen lamp would likely last for at least two to three years (depending on the hours of use), extending into the early 2020s. Incenting a customer to instead install an LED lamp would thus achieve savings into the early 2020s.

This approach (or a modified approach that effectively derives the same sunset-year outcome) is currently being used in a number of states, such as:

- **Massachusetts**: Uses a market adoption model to model decreasing savings over time. For Program Year 2016, this effectively allowed savings through 2021 for lamps subject to the 2012–2014 EISA GSL requirements, and 2022 for previously exempt lamps that are subject to the EISA requirements in 2020.
- **Arkansas**: Allows savings to be claimed through 2022 for all CFLs and LEDs.

4.4 Value Line LEDs

As LEDs have begun replacing CFLs in energy efficiency programs, the vast majority of program administrators have incented ENERGY STAR LEDs and have chosen not to include non-ENERGY STAR—referred to as "value line"—LEDs in their programs. Value line LEDs are defined as non-ENERGY STAR bulbs that are discounted well below the price of ENERGY STAR LEDs, are often in-house retailer generic-branded bulbs, and have a lower rated lifespan than ENERGY STAR bulbs. This is typically in response to some of the earlier quality challenges with CFLs and concern that if customers have a negative experience (due to poor quality or shorter-than-expected lifetimes) as they first try and then increasingly adopt LEDs, that this could lead to backsliding and negative impressions of the burgeoning technology.

In assessing the delta watts, however, value line LEDs pose a potential challenge because they typically offer a nearly identical wattage as the ENERGY STAR-equivalent lamps. The savings are the same; however, the difference in lifetime can lead to cost savings or other benefits, and dealing with that in detail is complex and common current methods often simply treat them as having the same savings, at both the net and gross level. The question arises: If a program is responsible for shifting customers from a non-ENERGY STAR to an ENERGY STAR LED, should there be any first-year savings?

This protocol recommends evaluators address the shift of sales from non-ENERGY STAR to ENERGY STAR lamps through the estimates of net-to-gross (NTG).[16] At the time of revision to this protocol, most methods of assessing lighting NTG (i.e., intercept surveys, elasticity modeling, sales data modeling, supplier interviews) do not differentiate between value line and ENERGY STAR lamps—that is, the baseline, or counterfactual condition, is assessing the total estimated sales of LEDs in the absence of program intervention. This means that if the baseline/counterfactual condition includes value line LEDs, the estimated "lift" due to program attribution is effectively capturing only the increased sales due to the program above the baseline sales of value line LEDs. In turn, the net savings are already being discounted for the presence and likely sale of value line LEDs.[17]

4.5 Annual Operating Hours

Hours of use (HOU) represents the estimated hours per year that consumers will use the energy-efficient lighting product. Metering studies have shown that the estimated average HOU for efficient lighting ranges from a low of 1.5 hours to a high of 3 hours per day (see Table 5), and have also demonstrated that self-reporting is not accurate. Myriad factors affect the expected number of hours per year that consumers use energy-efficient lighting products, including differences in demographics, housing types and vintages, efficient lighting saturation, room type, electricity pricing, annual days of sunshine, and even an "urban canyon" effect. Thus, data

---

[16] For jurisdictions that do not adjust savings for NTG, savings cannot be similarly adjusted for the shift from value line to ENERGY STAR LEDs.

[17] This approach may not account for other potential benefits of ENERGY STAR LEDs over value line LEDs, the most significant of which is likely longer lifetimes. When using NTG as an approach to incorporate this sales shift, the lifetime net benefits may be conservative/understated. To account for this, evaluation, measurement, and verification needs to specifically identify the percentage of program participants who shifted from value line to ENERGY STAR LEDs, then make assumptions about their net lifetime benefits.

extrapolation from one region to another has not successfully accounted for these influencing factors (Navigant Consulting and Cadmus 2011).[18] If extrapolation must be done (because a program was recently launched or because insufficient resources are available to conduct a metering study), evaluators may use secondary data from other metering studies (discussed in greater detail in Section 4.10). Based on these disparate results, this protocol recommends that program administrators—either on their own or through collaborations with neighboring utilities—collect primary data through a metering study of residential lighting measures.

Table 5. Estimated Efficient Lighting HOU From Recent Metering Studies

Region	Author	Sample Size (Homes)	# of Efficient Bulbs Metered	Estimated Average Daily HOU	Inclusive of LEDs
Maine	NMR Group, Nexant (2015)	67	488	2.0	Yes
Pennsylvania (All EDCs)	GDS, Nexant, RIA, Apex Analytics (2014)	216	518	3.0	Yes
California (PG&E, SCE, and SDG&E service areas)	KEMA, Inc. and Cadmus (2010), DNV GL (2014)	≈1,200	N/A	1.9 (2006-2008 cycle) 1.7 (2010-2012 cycle)	No
Georgia (Georgia Power Company)	Nexant and Apex Analytics LLC (2013)	125	594	2.8	No
Massachusetts, Rhode Island, Vermont, Connecticut	Nexus Market Research, Inc. et al. (2009)	157	657	2.8	No
Massachusetts, Rhode Island, Connecticut, New York	NMR Group and DNV GL (2014)	848	5,730*	2.9 (efficient bulbs) 2.7 (all bulbs)	Yes
Michigan	Opinion Dynamics and Cadmus (2012)	153	710	2.26	No
Illinois	Navigant Consulting (2012)	67	527	2.7	No
North Carolina (Duke Energy Progress)	Navigant Consulting, Apex Analytics, LLC (2012)	100	413	2.9	No
Maryland (EmPOWER)	Cadmus and Navigant Consulting (2011)	61	222	3.0	No
North Carolina, South Carolina	TecMarket Works and Building Metrics (2011)	34	156	2.5 (NC) 2.7 (SC)	No
Ohio	Vermont Energy Investment Corporation (from Duke Energy)	N/A	N/A	2.8	No
Pacific Northwest	Northwest Regional Technical Forum, based on California, KEMA, Inc., and Cadmus (2010)	N/A	N/A	1.9 (existing homes), 1.5 (new homes)	No

*Indicates count for both efficient and inefficient bulbs metered.

[18] This study revealed a significant difference in average daily HOU compared to extrapolating the HOU from the ANCOVA model (KEMA and Cadmus, 2010).

Although primary data collection through a metering study of residential lighting measures is the preferred approach, these analyses are usually limited to estimating operating hours for efficient versus inefficient lighting types. With the advent of LEDs into program administrator offerings, there has been interest in attempting to estimate the annual operating hours for both CFLs and LEDs. Unfortunately, most program administrators do not have the budget and resources required to meet statistical significance by sampling for each efficient bulb type. Therefore, one approach some jurisdictions have used to estimate CFL versus LED operating hours is to develop room-based annual operating hours and use the room-based saturations (NMR Group 2016). Overall bulb-type weighted operating hours can be estimated by evaluating the operating hours by each room's bulb-specific saturations. Until such time as administrators have the resources to meter at the specific bulb-type level, this protocol recommends using the room-based saturation approach as the best alternative.

4.6 Peak Coincidence Factor

Peak coincidence factor is typically defined as the fraction of the peak demand of a population that is in operation at the time of system peak. Thus, it is the ratio of the population's demand at the time of the system peak to its demand at the time of its own peak. For residential lighting, it represents the amount of time lights are on during the peak period, divided by the total time in the peak defined period (that is, the percentage of time that lights are on during the peak period). Note that although the methods below focus on HOU, the same principles apply to the estimate of the peak coincidence factor (including using the room-based saturations to develop bulb-specific peak coincidence factors as noted above for annual operating hours). For more information on the definition of peak demand, see the UMP *Chapter 10: Peak Demand and Time-Differentiated Energy Savings Cross-Cutting Protocols*.

4.7 Metered Data Collection Method

The metering approach needs to specify and manage the following factors and associated guidelines:

- Logger type
- Length of metering period
- Information collected on-site
- Data integrity.

4.7.1 Logger Type

This protocol recommends change-of-state loggers over periodic readings for standard bulbs as they can capture short intervals and switch rates (the number of times lights are turned on and off). For dimmable and three-way bulbs, the protocol recommends using light-intensity loggers. Current-sensing meters (rather than light-sensing meters) are an effective approach for outdoor conditions where ambient light can potentially inflate the estimated HOU.

4.7.2 Length of Metering Period

The length of the metering period depends on the focus of and available resources for the study. For example:

- If the intent of the study is to measure energy usage *without* concern for estimating summer peak demand (coincidence factor), use a limited metering period. Evaluators can limit the metering period to several weeks before and after the equinox (spring or fall). The general premise supporting annualizing metering periods shorter than one year is that the annual average use occurs precisely on the equinox; in fact, the equinox represents the annualization equation's intercept. A 2013 study demonstrated the precision of relying on a short period surrounding the equinox relative to using a complete 12 months of metering data (Shepherd et al. 2013).
- If the metering study in question is concerned *with both energy and demand*, conduct logging for at least six months and capture summer, winter, and at least one shoulder season (fall or spring). Ideally, evaluators should install loggers immediately preceding either the summer or winter solstice to capture a complete six months of data. In this case, an annualization adjustment is not required. If the metering period is shorter than six months *and* the meter placement is not coincident with the solstice, annualize the data—using techniques such as sinusoidal modeling—to reflect a full year of usage (DNV GL 2014).[19]

4.7.3 Information Collected On-Site

Conduct a complete inventory of lighting at all homes participating in the metering study. To allow for an estimate of saturation of high-efficiency lighting, the auditors should record the number and types of high-efficiency lighting by fixture and room type, and conduct a full inventory of sockets. Evaluators should collect on-site information specifically related to the logger placements that details room type, window orientation, fixture type, notes about possible ambient light issues, etc.

4.7.4 Data Integrity

Clean and thoroughly check all metered data for errant and erroneous observations. For example, at the moments of installation and removal, clip the downloaded data to eliminate extraneous readings. Also, omit data from broken loggers or loggers removed by residents. Also omit data from loggers suspected to have metered daylight/ambient light. Finally, examine the data for "flicker" (that is, very frequent on/off cycling) and clean the raw data to correct for flicker. Evaluators can perform computer programming via R, SAS, or other statistical software that allows data from flickering bulbs to effectively remain on for the duration of the flickering event, rather than appear to be repeated on/off events.

4.7.5 Metering Sample Design

Ideally, evaluators should conduct metering for large samples of all major lighting types (including incandescent or halogen baseline bulbs); however, in practice, most evaluators do not have adequate resources for a scope of this size. Consequently, to optimize the allocation of moderate evaluation resources, it is important to target the metering to select lighting measures—typically CFLs or LEDs—that represent the greatest savings in a residential lighting program. (This is especially true for retrospective program savings). Where savings are used prospectively,

---

[19] Sinusoidal modeling assumes that HOU will vary inversely with hours of daylight over the course of a year. Sinusoid modeling shows that HOU change by season, reflective of changes in the number of daylight hours and weather, and that these patterns will be consistent year to year, in the pattern of a sine wave.

it is important to attempt to meter all lighting types, as studies have found that efficient bulbs tend to be installed in higher-use fixtures first and therefore have higher HOU than average bulb types (KEMA and Cadmus, 2010, NMR Group and DNV GL 2014)).[20]

Given the difficulty of identifying program bulbs in an upstream program, field technicians may place loggers on energy-efficient bulbs in a random sample of homes that have installed similar measures, even if those measures are not definitely known to be part of a markdown or buy-down program. For homes that have many energy-efficient lighting products, evaluators may meter a subsample of bulbs, if they are selected randomly within the home. For example, if a home has LEDs in 10 fixtures, place meters on three to five randomly selected fixtures (DNV GL 2014).[21] This placement will minimize the invasiveness in homes that are highly saturated with energy-efficient lighting products and be cost effective, enabling metering of a larger sample of bulbs in an equivalent number of homes.

Understanding that any metering study is likely constrained by resource and budget limitations, as noted above, evaluators should set expectations for the desired levels of statistical confidence and precision based on the likely number of meters deployed in the field, and assume a coefficient of variation (CV) based on recent studies of programs with similar CFL or LED saturation (using the maturity of program as a proxy, if necessary) and housing characteristics (Cadmus 2010; Navigant Consulting and Cadmus 2011). Historically, the CV has been assumed (and sometimes reported) as approximately 0.5 or 0.6. However, this CV may be considerably too low when accounting for the serial correlation of usage (and error) across light circuits within a home. For example, a recent lighting HOU study from New England, based on more than 800 homes and 5,700 loggers, recommends that evaluators use a CV of 1.2 for all rooms combined, with CVs ranging from 0.89 to as high as 1.6 by room type, as shown in Table 6 (NMR Group and DNV GL 2014).

Table 6. Example of Calculated CV from a Lighting Metering Study

Room Type	CV	Sample Size for 90/10 Confidence/Precision (# of Loggers)
Bathroom	1.38	515
Bedroom	1.15	358
Dining Room	1.10	327
Exterior	0.89	214
Kitchen	0.93	233
Living Space	1.04	293
Other	1.60	693
Household	1.20	390

This protocol recommends that, at a minimum, room type be considered as a within-home

[20] For example, the NMR metering study from New England estimated daily HOU for all bulbs at 2.7 hours/day, but 3.0 hours/day for efficient bulbs. However, the authors do not believe this difference is due to saturation, but rather to a combination of selective installation (that is, higher use sockets and fixture use) and potentially to snapback (see Section 4.9).

[21] A number of studies, including the evaluation of the 2006–2008 California Upstream Lighting Program, provide publicly available examples of how to randomly select fixtures for metering.

sampling stratum as room type is one of the most important determinants of HOU. Therefore, the program administrator should work with the program evaluator to establish well-defined targets for the total number of room types to meter. Stratifying by room types (rather than by home type) allows for a potentially more homogeneous population unit because of more consistent usage within room types. It is also important to estimate the HOU by room type because direct-install programs often target higher-use fixtures and sockets in higher-use rooms. If program administrators track the room types associated with the installation of efficient lighting products, evaluators can then base HOU on room type.

When calculating the HOU from the meter data collected, the precision estimates should take into account the primary sampling unit (household) and other subsample units (room type). Most statistical packages used for HOU estimation allow for clustering of the sampling unit (household) to account for correlation.

The confidence and precision of the HOU estimate is not simply a factor of the variance across each hour for each logger. Using these units would lead to grossly overestimated precision (i.e., appears more precise) if based on every hour across the metering period. Furthermore, the evaluator's calculations should not ignore the error inherent in the HOU from an annualization model. Rather, when estimating the overall HOU, any evaluator's model or calculation should estimate the annualized HOU for each logger across all hours and treat this as one observation, account for the error across all loggers, and then use these estimates as the starting point for the room- and household-based averages.

Following the metering effort and the annualization of results, weight the HOU to reflect the actual distribution of lighting products by room type. For example, if 10% of the loggers are installed in kitchen fixtures, but the audit data reveal that 15% of all CFLs are installed in kitchens, weight the data from the loggers in kitchens up by 1.5 when calculating total HOU.[22]

It is also important to estimate the HOU by room type because direct-install programs often target higher-use fixtures and sockets in higher-use rooms. If administrators of these programs track the room types associated with the installation of efficient lighting products, evaluators can then base HOU on room type.

Evaluators should also compare the demographic and household characteristics of the metering sample with the characteristics of the total population of households believed to have purchased energy-efficient lighting products. (Evaluators can collect this information through telephone or web-based surveys.) If significant differences appear *and* there is a large enough sample to support reweighting based on such characteristics, evaluators should weight the results to reflect these differences.

4.8 Using Secondary Data

Metering is the recommended approach; however, program administrators who are just launching a program—or who do not have sufficient resources to conduct a metering study—

[22] If there are differential sampling rates within a room type, the sampling rates also need to be accounted for in the weighting.

may use secondary data from other metering studies. This protocol recommends using the following criteria when selecting and using secondary data to estimate HOU:

- Similarities in service territories
- Adequate sample size for reasonable confidence/precision levels
- Length of metering period
- Adjustments to reflect hours of use by room type.

4.8.1 Similarities in Service Territories

Selecting a similar service territory based on geographic proximity or latitude and as many common demographic and household characteristics as possible will increase the likelihood that the secondary data will provide a valid, reasonable, and accurate estimate.

4.8.2 Sample Size

The number of observations varies considerably between studies, so evaluators should compare the sample size, standard errors, and precision levels at equivalent confidence levels across studies to ensure a selected study has reasonable confidence and precision levels.

4.8.3 Length of Metering Period

The protocol recommends selecting studies that capture both winter and summer usage to estimate summer and winter peak demand, when demand is a critical factor, or may select studies that captured usage over a shorter period when energy is the only variable of interest (see Section 4.8.2 above).

4.8.4 Adjustments to Reflect Hours of Use by Room Type

To extrapolate HOU from one region to another, one approach is to calibrate the HOU based on the efficient bulb saturation by room type. If possible, weight the HOU by room type from a secondary data source by the room type distribution of efficient lighting for the region under study.

4.9 Snapback/Rebound or Conservation Effect

Snapback or **rebound** refers to changes in use patterns that occur after an energy-efficient product is installed, resulting in reducing the overall measure savings. For example, when residential lighting customers use a CFL or LED for more hours per day than they had used the replaced incandescent bulb, without a corresponding reduction in use of another less efficient lamp, this constitutes snapback. This behavior change may be because of factors such as the cost savings per unit of time from the CFL or LED or a concern that turning CFLs or LEDs on and off shortens their effective useful life (although most consumers are probably unaware of this effect). Some customers, however, might have lower HOU after installing a CFL or LED because they also want to reduce energy consumption or are dissatisfied with the quality of light.

Residential lighting programs do not typically allow metering to be conducted both before and after the installation of energy-efficient lighting. However, a recent lighting study in the Northeast found that the HOU were higher for sockets with efficient bulbs compared to all sockets in the house (NMR Group and DNV GL 2014). The difference was believed to be for these three reasons: 1) differential socket selection (households selecting higher use locations for their high-efficiency light bulbs), 2) shifting usage (households install an efficient bulb in a socket and then begin to use that socket in lieu of sockets containing inefficient bulbs), and/or 3) snapback. However, this evaluation did not collect any data to determine which of these three theories is correct or the proportion of the difference between efficient and inefficient HOU attributable to each type of behavior. Unfortunately, this protocol cannot recommend researching for snapback/rebound effects as there is currently no way to estimate other than the highly unreliable self-report approach.

4.10 In-Service Rate

The ISR represents the percentage of incented residential lighting products that are ultimately installed by program participants. ISRs vary substantially based on the program delivery mechanism, but they are particularly important in giveaway or upstream programs where the customer is responsible for installation.

For the upstream programs shown in Table 7, three factors have led to first-year ISRs (LEDs installed within the first year after acquisition) below 100%:

- Often deeply discounted prices
- Inclusion of program multipacks
- Consumers commonly waiting until a bulb burns out before replacing it.

Table 7. Estimated First-Year ISRs from Recent Evaluations of LED Upstream Lighting Programs

Region	Author	Percentage of LEDs Installed the First Year After Purchase*
Massachusetts	NMR Group, Inc. (2016)	84%
Connecticut	NMR Group, Inc. (2016)	95%
Colorado	Cadmus (2016)	84%
Maine	NMR Group, Inc., and Nexant (2016)	94% (phone)
Wisconsin	Apex Analytics and Cadmus (2016)	99%

*Based on program year only, not years subsequent to the program year or several years in a multiyear program cycle.

This protocol recommends that evaluators use the methods appropriate to the specific delivery mechanism to estimate ISRs:

- For **direct-install programs**, conduct verification (such as telephone survey or site visits) to assess installation and early removal (that is, removal prior to failure).
- For **giveaway or coupon programs**, conduct verification when customer contact information is available. Also, ask respondents whether the installation location was within the relevant service territory and whether the measure was installed in a home or a business. If the installation was in a business, ask about the type of business.

- If customer information is not available, rely on either secondary data (such as those from a similar program where customer information was collected) or on the in-home audit approach (described in the next bullet).
- For **upstream programs**, calculate ISRs through in-home audits. Because program bulbs cannot be easily identified, evaluators can calculate the ISR as the number of installed bulbs purchased in a recent 12-month period divided by the total number of bulbs purchased in the same 12-month period. If the sample size of homes with bulbs purchased in a recent 12-month period is insufficient to provide the necessary levels of confidence and precision, apply a long-term ISR using all bulbs, regardless of the time of purchase.
- Although the in-home audit is the recommended approach, evaluators can use a telephone survey when program administrators are just **launching a program or have insufficient resources** to conduct an in-home audit. To minimize recall bias, the callers should focus questions only on products purchased in a recent 12-month period rather than the period covering the long-term ISR. (Respondents are expected to have better recall about the percentage of bulbs purchased and installed within the past 12 months compared to the percentage of bulbs they have ever purchased and installed.)

Although first-year ISRs for upstream programs are less than 100%, recent studies have demonstrated that consumers plan to install most of the incented bulbs; however, consumers often wait until an existing bulb burns out (Navigant and Itron 2010).[23] As a result, for savings that occur in years following the year that the incentive was paid, program administrators have used the following approaches to account for bulbs that are subsequently installed:[24]

- **Stagger the timing of savings claims.** In this method, all the program expenses are claimed during the program year, but the savings (and therefore the accompanying avoided-cost benefits) are claimed in the years during which the program measures are estimated to be installed. This approach more accurately captures the anticipated timing and quantity for the realized savings.
- **Discount future savings.** In this method, all the costs and benefits are claimed during the program year, but the savings (in terms of avoided costs, kilowatt-hours, or kilowatts) from the expected future installation of stored program bulbs are discounted back to the program year using a societal or utility discount rate.[25] This method offers the simplicity of claiming all benefits and costs during the program year and thus not having to track and claim future installations.

To calculate the installation rate trajectories, this protocol recommends using the findings from a Massachusetts panel study (NMR Group 2017). The Massachusetts study included 105 homes

[23] For example, the evaluation in the Navigant and Itron study (2010) found that about 90% of customers were waiting until an incandescent or CFL burned out before they installed a stored CFL (Table 3 through Table 6).
[24] The selection of approach depends on the study's purpose and regulatory requirements.
[25] Energy or demand savings are not normally discounted; however, this approach provides simplicity for calculating program benefit/cost ratios and the actual net present value of avoided costs, which often are used for cost recovery. For programs that want to bid into capacity markets (for example, PJM), the staggered approach is recommended because it more accurately captures the actual timing and cumulatively increasing nature of the demand savings.

with 991 LEDs and looked at ISRs for up to two years (including lamps that were initially placed in storage the first year after purchase).

The Massachusetts study found that 24% of the LEDs that went into storage in year 1 were installed in year two. Although the study is expected to have a three-year ISR available in early 2018, only two years of data were available at the time this protocol was being revised. Therefore, to estimate the lifetime ISR, evaluators can assume customers continue to install LEDs in storage at a rate of 24% of stored bulbs each year.

Evaluators can follow this trajectory and calibrate to individual service territories using the example below. As outlined in Table 8, program administrators use their researched value for the Year 1 ISR and determine the percentage of stored bulbs installed in each of the next few years:

- Year 2 installation of stored bulbs is calculated by multiplying the percentage of bulbs in storage by 24% and adding that to the first-year ISR. In this example, 24% of the stored LEDs (24%*25%=6%) will be installed in Year 2, bringing the Year 2 ISR to 81%
- Year 3 installation of stored bulbs is calculated by multiplying the percent of bulbs still in storage after Year 2 by 24% (24%*19%=5%) and adding that to the second year ISR, bringing the cumulative ISR in this example up to 86%.

Table 8. Estimated Cumulative 3-Year ISR Calculations

Year	Cumulative ISR*	Cumulative Storage Rate
Year 1	75%	25%
Year 2	81%	19%
Year 3	86%	14%

*This rate represents the percentage of bulbs purchased in Year 1 and installed by the end of each following year. The first year ISR of 75% is only an example, and evaluators should use researched values for the first-year ISR.

However, it is recognized that bulbs may continue to be installed for multiple years and that estimating the lifetime ISR also requires consideration of the effective useful life (EUL) of the lamp. In the example above, a 2017 program would have 25% of the program LEDs initially go into storage but then would continue to have program-incented lamps installed into the early 2020s. As noted above in the lifetime savings discussion, however, programs may be truncating the EUL of LEDs to account for the EISA backstop provision.

This protocol, therefore, recommends also truncating the ISR trajectory year at the year in which the EUL of lamps is reduced. In other words, lamps installed after that year can no longer claim savings if the baseline becomes an efficient lamp. Using the example above, assume that Year 1 is 2017 and that the cumulative ISR was extended out five years (to 2021), which would increase the cumulative ISR to 92%. If 2021 is assumed to be the "sunset year" for claiming savings on LEDs, lamps installed after 2021 would not claim any additional savings, and thus the ISR would be capped at 92%. As noted above, however, if the future installations are claimed in the year in which the bulbs are incented, the future installations should be discounted back to the program year using a societal or utility discount rate. This could have the effect of decreasing the cumulative ISR.

4.11 Interactive Effects with Heating, Ventilating, and Air Conditioning

CFLs and LED lamps emit less waste heat than incandescent bulbs, which affects heating, ventilating, and air conditioning (HVAC) energy requirements. These effects vary based on space conditioning mode, saturation of space heating and cooling technologies and their relative efficiencies, and climate zones. The influence of climate zone on interactive effects depends on a variety of house-specific factors.

Taking all of these factors into account, the net impact on lighting energy cost savings could be positive, negative, or neutral (Parekh 2008, Parekh et al. 2005). In cooling-dominated climates, the interactive effects are positive, resulting in additional savings from decreased cooling load. However, in heating-dominated climates, the interactive effects are negative, with decreased savings from increased heating load.

Because of the potential impacts of interactive effects, the Residential Lighting Evaluation Protocol recommends including these effects in evaluations of residential lighting programs.[26] One common approach is to estimate these effects with simulation models, examining a mix of typical housing types (such as different vintages) and reflecting the estimated saturation, fuel shares, and size/efficiency of HVAC equipment. (That is, the percentage of homes that have air conditioning or electric versus gas heat.) If necessary, use secondary sources—such as the Residential Energy Consumption Study (U.S. Department of Energy 2015)—to estimate these inputs.

Some regions have developed interactive effects calculators based on such simulations (for example, in California, the Database for Energy Efficiency Resources and the Regional Technical Forum in the Northwest.[27] Such regional collaboration can minimize the cost of determining the interactive effects for regions that do not already have such a tool.

If regional collaboration is not an option *and* the program administrator does not have the resources to complete the simulations, the Residential Lighting Evaluation Protocol recommends using a value from an existing resource. It recommends that the value used reflects key similarities between the program administrator's territory and the territory from which the data are taken. At a minimum, these key similarities should be the climate (heating and cooling degree days and, ideally, the latitude), HVAC system types, HVAC fuel types, and HVAC system saturations.

[26] Interactive effects are relevant for bulbs installed in conditioned spaces only. Thus, exterior lights will not have HVAC interactive effects.

[27] http://www.deeresources.com/files/DEER2013codeUpdate/download/DEER2014-Lighting-IE_and_Adjustment-Factor-Tables-17Feb2014.xlsx

5 Other Evaluation Issues

The incentive structure of upstream lighting programs does not inherently allow for assurances that each purchaser of a program bulb is a residential customer in the sponsoring program administrator's service territory. Therefore, some program bulbs may be purchased by nonresidential customers or by customers served by other utilities. This section discusses these parameters; Section 5.3 provides recommended approaches.

5.1 Sales to Nonresidential Customers

Nonresidential customers typically use lighting products for more hours per day than residential customers. Typically, nonresidential customers also have higher peak coincidence factors. Therefore, the lighting products incentivized through a residential lighting program but installed in nonresidential sockets may lead to higher savings than those assumed through the previously discussed methods. A recent literature review of 23 cross-sector sales studies found that average cross-sector sales for upstream programs was approximately 7% (Cadmus 2015).

Evaluators estimate this parameter via several different methods, including:

- **Customer intercept surveys:** At the time of sale, customers who purchase lighting products participate in a short survey about intended installation location and facility type
- **Surveys with store managers:** Asking managers to estimate the percentage of bulbs sold to nonresidential customers
- **Residential customers:** Asking customers if they purchased discounted lighting products and installed them in businesses
- **Owners of small businesses:** Asking business owners where they typically purchase lighting products.

Key limitations in estimating this parameter are recognized in this protocol:

- **Customer intercepts may not represent all program sales.** Conducting customer intercept surveys can be expensive, and evaluators may conduct them only in high-volume stores (such as Home Depot, Lowe's, and Walmart) to minimize the cost per survey. In some cases, these surveys are conducted only during high-volume promotions. Also, because some retailers refuse to allow the surveys on their premises, the surveys may not be representative of total program sales.
- **Accuracy from intercepts is further challenged because business owners and contractors may be a minority of purchasers, leading to smaller respondents in the sample.** This challenge may be heightened because nonresidential customers may not purchase during the same timeframes as the average residential purchaser, and they may purchase significantly larger quantities (thus, a small number of respondents may skew the results).
- **Surveys lack high reliability.** Store managers usually do not have detailed information about program bulb purchasers, so their estimates of sales to nonresidential customers may be unreliable. There are also challenges when surveying small business customers, such as nonresponse bias (that is, calling a small business and not getting cooperation

from the business decision-maker to take a survey). Recall bias among survey participants may also make quantifying the number and type of bulbs acquired by nonresidential purchasers difficult.

5.2 Cross-Service Area Sales (Leakage)

Recent studies have also attempted to estimate the number of program bulbs sold to customers outside the program administrator's service territory. This is commonly referred to as *leakage* or *spillage*.

The most common approaches to determining leakage are clearly delineated in the Arkansas TRM (Arkansas Public Service Commission 2016),[28] cited below in per order of that TRM.

- **Customer intercept surveys.** This is the preferred method of primary data collection for actual participants, although it can be very difficult to receive permission from participating retailers. The sampling strategy used should attempt a random mix of entities (geographic, retailer, day of week, and avoiding promotional events only).
- **Geo-mapping with general population surveys.** This method involves modeling leakage scores based on the geographic proximity of participating retailers to sponsoring utility customers relative to other utility customers (non-sponsoring). Evaluators can refine the model by using general population telephone surveys to confirm purchasing behavior for sponsoring and non-sponsoring utilities in the region.
- **Opt-in surveys.** This involves including a label or note with each incented product among all participating retailers with instructions about how to participate in survey. (Ideally, the survey should be multimodal: reply card, online, and phone number.) Low response rates and nonresponse bias are drawbacks.

Estimated leakage could vary substantially based on the service territory and program design, with recent estimates as high as 65.4% (Cadmus 2014), and as low as 2.1% (Illinois Energy Efficiency Stakeholder Advisory Group 2017) or < 1% (KEMA and DNV-GL 2014).

Key limitations in estimating leakage are recognized in this protocol:

- **Cross-region sales.** Many neighboring service territories are now targeted by residential lighting programs; thus, there is a lesser incentive to shop outside one's own service territory to purchase less expensive lighting products. In some cases, program bulbs cross over in both directions across service boundaries, which may offset the effect in either or both territories.
- **Many programs now limit the number of participating retailers, so leakage is minimized**. Many program administrators now require retailers participating in upstream programs to be located far enough within the service territory or to be surrounded by a certain percentage of program customers to minimize potential leakage.

[28] See "Protocol K: Leakage" in the Arkansas 2013 TRM.

5.3 Estimating Cross-Customer Class and Cross-Service Area Sales

In addition to the limitations presented above, these parameters may also at least partially offset each other. That is, the increased savings of sales to nonresidential customers may be at least partially offset by leakage.[29] Given this, it is reasonable to exclude these parameter estimates from impact evaluations of upstream residential lighting programs. In addition, given the opposing directions of these parameters, either both—or none—of these parameters should be incorporated. Thus, do not claim increased savings from sales to business customers without also adjusting for leakage, and do not decrement program savings from leakage without also incorporating sales to business customers.[30]

[29] These protocols do not imply that these effects will offset exactly, only that they work in opposite directions; sales to nonresidential customers will typically lead to greater savings, and cross-service area sales will lead to lower savings in the sponsor's service territory. Note also that the longer HOU for commercial installations may, in fact, more than offset reduced savings from leakage. For example, if nonresidential HOU were shown to be four times the residential HOU in a given jurisdiction, a rate of 5% nonresidential installations would have an amplified effect of generating close to 20% of the overall energy savings for the program.

[30] Exceptions can be made in cases where program administrators are surrounded by other service territories offering similar programs. In these cases, sales to business customers can be claimed without reducing sales from leakage. An example of this is in Pennsylvania where the Phase II Evaluation Framework recommends that evaluation contractors assume that leakages into and out of each utility territory effectively offset each other because they offer the same or similar upstream lighting programs (Pennsylvania Public Utilities Commission 2013).

6 Looking Forward

Residential lighting programs offer a range of measures through multiple delivery strategies; the upstream LED program is currently the most ubiquitous. Program administrators who offer a variety of measures and rely on multiple delivery strategies may need to prioritize their evaluation resources based on criteria such as contribution to savings and assessed uncertainty. Evaluators should assess savings through a mix of primary and secondary data, using International Performance Measurement and Verification Protocol Option A (Retrofit Isolation: Key Parameter Estimates).

A key area that needs additional research involves the assumptions about the EISA backstop provision. This is an evolving area, with tremendous uncertainty over the regulations. Even if the regulations were to be fully repealed, however, LEDs have gained a tremendous amount of momentum and increasing market share, while future CFL production is increasingly unlikely. These factors should be considered when estimating lifetime savings for current programs.

In addition, the lifetime ISR trajectory is based only on a single panel study that offered two years of ISRs. Additional primary data on the ISR trajectory will be helpful to test the assumptions of a fixed trajectory beyond the second year.

7 References

Arkansas Public Service Commission (August 31, 2016). *Arkansas Technical Reference Manual Version 6.0*, pp. 204-205, Tables 192, 193. Available online: http://www.apscservices.info/EEInfo/TRM6.pdf

Cadmus (January 14, 2016). *Colorado Lighting Market Study*. Prepared for Xcel Energy. https://www.xcelenergy.com/staticfiles/xe/PDF/Regulatory/CO-DSM/CO-Regulatory-DSM-Lighting-Market-Study.pdf

Cadmus (March 2015). *Massachusetts Residential Lighting Cross-Sector Sales Research.* Prepared for the Massachusetts Program Administrators and Energy Efficiency Advisory Council.

Cadmus (March 12, 2014). *SWEPCO Energy-Efficiency Portfolio Evaluation Report 2013 Program Year*. Prepared for SWEPCO. http://www.apscservices.info/EEInfo/EEReports/SWEPCO%202013.pdf

Cadmus and NMR Group, Inc. (June 8, 2013). *Residential Lighting Shelf Survey and Pricing Analysis, Final Report.* Prepared for Cape Light Compact, National Grid, NSTAR Electric, Unitil, Western Massachusetts Electric Company, and Energy Efficiency Advisory Council Consultants.

Cadmus (December 2010). *Lighting and Appliance Evaluation—PY 2*. Prepared for Ameren Illinois.

DNV GL (August 4, 2014). *California Upstream and Residential Lighting Impact Evaluation: Work Order 28 (WO28) Final Report*. Prepared for California Public Utilities Commission, Energy Division.

ENERGY STAR. "Certified Products." Available online: https://www.energystar.gov/products/specs/lamps_specification_version_1_0_pd

ENERGY STAR Program Requirements for Lamps (Light Bulbs) (October 2016). Available online: https://www.energystar.gov/sites/default/files/ENERGY%20STAR%20Lamps%20V2_0%20Revised%20OCT-2016_1.pdf

Federal Trade Commission (May 2013). "The FTC 'Lighting Facts' Label: Questions and Answers for Manufacturers." Available online: http://www.business.ftc.gov/documents/bus26-lighting-facts-questions-and-answers-manufacturers

GDS Associates Inc.; (January 13, 2014). Pennsylvania Statewide Act 129 2014 Commercial and Residential Lighting Metering Study. Prepared for the Pennsylvania Public Utilities Commission by the Statewide Evaluator Team (GDS Associates and Nexant)

Illinois Energy Efficiency Stakeholder Advisory Group (February 8, 2017). *IL Statewide Technical Reference Manual, Version 6.0.* Page 243-245. Available online: http://www.ilsag.info/il_trm_version_6.html.

KEMA, Inc., and DNV-GL (August 4, 2014). *California Upstream and Residential Lighting Impact Evaluation*. Prepared for California Public Utilities Commission, Energy Division. http://www.calmac.org/publications/WO28_-_California_Upstream_and_Residential_Lighting_Impact_Evaluation_Final_Report.pdf

KEMA, Inc., and Cadmus (February 8, 2010). *Final Evaluation Report: Upstream Lighting Program*. Vol. 2. Prepared for California Public Utilities Commission, Energy Division. http://www.calmac.org/publications/FinalUpstreamLightingEvaluationReport_Vol2_CALMAC.pdf

KEMA, Inc. (February 25, 2005). *CFL Metering Study*. Prepared for California's Investor-Owned Utilities (Pacific Gas & Electric Company [PG&E], Southern California Edison [SCE], and San Diego Gas and Electric [SDG&E]).

Lightopedia. *United States Congress. Energy Independence and Security Act of 2006.* H.R.6. Available online: http://www.lightopedia.com/_files/eisa/energy-independence-and-security-act-of-2007.pdf

Navigant Consulting and Apex Analytics, LLC (November 7, 2013). *Storage Log Study of CFL Installation Rate Trajectory.* Prepared for Duke Energy Progress.

Navigant Consulting, Itron, Opinion Dynamics, and Michaels Energy (October 23, 2013). *Residential ENERGY STAR Lighting PY5 Evaluation Report.* Presented to Commonwealth Edison Company. http://www.ilsag.info/comed_eval_reports.html

Navigant Consulting and Apex Analytics, LLC (September 5, 2012). *Preliminary Shelf-Stocking Study Results for the Residential ENERGY STAR Lighting Program.* Prepared for Progress Energy Carolinas.

Navigant Consulting, Cadmus, and Apex Analytics LLC (June 27, 2012). *EM&V Report for the 2010-2011 Residential ENERGY STAR Lighting Program*. Prepared for Duke/Progress Energy Carolinas.

Navigant Consulting and Cadmus (January 15, 2011). *EmPOWER Maryland 2010 Interim Evaluation Report*. Prepared for Baltimore Gas and Electric, Potomac Electric Power Company, Delmarva Power and Light (DPL), Southern Maryland Electric Cooperative (SMECO), and Allegheny Power (AP).

Navigant Consulting and Itron (December 21, 2010). *Plan Year 2 Evaluation Report: Residential ENERGY STAR Lighting*. Prepared for Commonwealth Edison Company.

Nexant, Inc. and Apex Analytics LLC (October 4, 2013). *Georgia Power Company Residential Lighting Saturation and Metering Study*. Prepared for Georgia Power Company.

Nexus Market Research, Inc.; RLW Analytics, Inc., and GDS Associates (January 20, 2009). *Residential Lighting Markdown Impact Evaluation*. Prepared for Markdown and Buydown Program Sponsors in Connecticut, Massachusetts, Rhode Island, and Vermont. http://www.env.state.ma.us/dpu/docs/electric/09-64/12409nstrd2ae.pdf.

NMR Group (April 7, 2017). *RLPNC 16-7: 2016-17 Lighting Market Assessment Consumer Survey and On-site Saturation Study.* Prepared for The Electric and Gas Program Administrators of Massachusetts. http://ma-eeac.org/wordpress/wp-content/uploads/Lighting-Market-Assessment-Consumer-Survey-and-On-Site-Saturation-Study.pdf

NMR Group (November 7, 2016). *Residential Lighting Hours-of-Use Update.* Memorandum to Massachusetts Electric PAs and EEAC Consultants. Available online: http://ma-eeac.org/wordpress/wp-content/uploads/Residential-Lighting-Hours-of-Use-Update.pdf

NMR Group (January 28, 2016). *Connecticut LED Lighting Study Report (R154).* Prepared for Connecticut Energy Efficiency Board (EEB). https://www.energizect.com/sites/default/files/R154%20-%20CT%20LED%20Lighting%20Study_Final%20Report_1.28.16.pdf

NMR Group and Nexant (April 16, 2015). *Efficiency Maine Retail Lighting Program Overall Evaluation Report.* Prepared for Efficiency Maine. http://www.efficiencymaine.com/docs/Efficiency-Maine-Retail-Lighting-Program-Evaluation-Report-2015.pdf

NMR Group and Apex Analytics, LLC (May 22, 2014). *Market Effects, Market Assessment, Process and Impact Evaluation of the NYSERDA Statewide Residential Point-of-Sale Lighting Program: 2010-2012.* Prepared for NYSERDA. https://www.nyserda.ny.gov/-/media/Files Publications/PPSER/Program-Evaluation/2014ContractorReports/2014-EMEP-pos-lighting-program.pdf

NMR Group and DNV GL (May 5, 2014). *Northeast Residential Lighting Hours-of-Use Study*. Prepared for Connecticut Energy Efficiency Board, Cape Light Compact, Massachusetts Energy Efficiency Advisory Council, National Grid Massachusetts, National Grid Rhode Island, New York State Energy Research and Development Authority.

Opinion Dynamics and Cadmus (August 7, 2012). *Michigan CFL Hours of Use Study.*

Parekh, A. (2008). "Do CFLs Save Whole-House Energy?" *Home Energy Magazine*, November/December 2008. pp. 20–22.

Parekh, A., Swinton, M.C., Szadkowski, F., and Manning, M. (2005). *Benchmarking of Energy Savings Associated with Energy Efficient Lighting in Houses*. National Research Council Canada. NRCC-50874. http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=shwart&index=an&req=20377557

Shepherd, B., Rambo, E., and Busker, M. (2013). *Annualization of Results of Residential Lighting Meter Data*, presented at 2013 International Energy Program Evaluation Conference, August 2013, Chicago, IL.

TecMarket Works; Building Metrics (February 15, 2011). *Duke Energy Residential Smart $aver CFL Program in North Carolina and South Carolina: Results of a Process and Impact Evaluation*.

U.S. Department of Energy (December 29, 2016). *Energy Conservation Program: Energy Conservation Standards for General Service Lamps*. Definition. 10 CFR Part 430. Docket Number EERE-2013-BT-STD-0051. RIN 1904-AD09. Available online: https://energy.gov/sites/prod/files/2017/01/f34/gsl_definition_finalrule_2016-12-29_0.pdf

Vermont Energy Investment Corporation (July 2011). *Mid-Atlantic Technical Reference Manual, Version 2.0*. Facilitated and Managed by Northeast Energy Efficiency Partnerships. http://neep.org/uploads/EMV%20Forum/EMV%20Products/A5_Mid_Atlantic_TRM_V2_FINAL.pdf

Vermont Energy Investment Corporation (August 6, 2010). *State of Ohio Energy Efficiency Technical Reference Manual*. Prepared for the Public Utilities Commission of Ohio. http://amppartners.org/pdf/TRM_Appendix_E_2011.pdf

8 Resources

American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) (2000). *Research Project 1093 Compilation of Diversity Factors and Schedules for Energy and Cooling Load Calculations.* ASHRAE Research Report.

ASHRAE (2002). *Guideline 14-2002: Measurement of Energy and Demand Savings*.

ASHRAE (2010). *Performance Measurement Protocols for Commercial Buildings*.

ASHRAE. (TBD). *Guideline 14-2002R* (Revision of *Guideline 14*, currently in process, publication date TBD).

Glacier Consulting (March 6, 2008). *Analysis of Delta Watts Values for CFLs Rewarded Through the Residential Lighting Program During FY07*. Prepared for the State of Wisconsin Public Service Commission. Available online: http://www.cee1.org/eval/db_pdf/1215.pdf.

National Renewable Energy Laboratory (March 2002). *International Performance Measurement & Verification Protocol: Concepts and Options for Determining Energy and Water Savings. Volume I.* Available online: http://www.nrel.gov/docs/fy02osti/31505.pdf

Pennsylvania Public Utilities Commission (May 31, 2013). *Evaluation Framework for Pennsylvania Act 129 Phase II Energy Efficiency and Conservation Programs.* Prepared by the Statewide Evaluation Team: GDS Associates, Nexant, Research Into Action, and Apex Analytics.

PA Consulting Group and NMR Group (April 22, 2010). *Focus on Energy Evaluation 2009 CFL Savings Analysis*.

RLW Analytics, Inc., Nexus Market Research (April 22, 2005). *Extended Residential Logging Results*. Prepared for National Grid.

Chapter 7: Refrigerator Recycling Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Josh Keeling
Cadmus
Portland, Oregon

Doug Bruchs
Cadeo
Portland, Oregon

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68563
September 2017

Contract No. DE-AC36-08GO28308

Chapter 7: Refrigerator Recycling Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Josh Keeling
Cadmus
Portland, Oregon

Doug Bruchs
Cadeo
Portland, Oregon

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68563
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

This chapter was originally developed in 2013 by Doug Bruchs and Josh Keeling of The Cadmus Group, Inc. The 2017 revision, led by Doug Bruchs (Cadeo) and Jason Christensen (Cadmus), benefitted from the input of multiple experts and stakeholders. The authors wish to thank members of the UMP Technical Advisory Group and acknowledge Sami Khawaja (Cadmus) and Peter Franzese (California Public Utilities Commission) for their thoughtful contributions.

Suggested Citation

Keeling, J.; Bruchs, D. (2017). *Chapter 7: Refrigerator Recycling Evaluation Protocol. The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68563. http://www.nrel.gov/docs/fy17osti/68563.pdf

Acronyms

CDD	cooling degree day
DOE	U.S. Department of Energy
EUL	effective useful life
HDD	heating degree day
kWh	kilowatt hour
NREL	National Renewable Energy Laboratory
NTG	net-to-gross
RUL	remaining useful life
UEC	unit energy consumption
UMP	Uniform Methods Project

Protocol Updates

The original protocol was published in April 2013.

This chapter has been updated to incorporate the following revisions:

- Added a unit energy consumption (UEC) regression model for freezers based on available *in situ* freezer-specific metering data
- Dropped the induced replacement adjustment as part of the net savings calculation due to the difficulty measuring the adjustment and its small impact on savings
- Highlighted the usefulness of recycling programs in cross-promoting efficiency programs and providing opportunities for positive customer engagement.

Table of Contents

List of Figures

List of Tables

1 Measure Description

Refrigerator recycling programs are designed to save energy by removing operable, albeit less efficient, refrigerators from service. By offering free pickup, providing incentives, and disseminating information about the operating cost of less efficient refrigerators, these programs are designed to encourage consumers to:

- Limit the use of secondary refrigerators[1]
- Relinquish refrigerators previously used as primary units when they are replaced (rather than keeping the existing refrigerator as a secondary unit)
- Prevent the continued use of less efficient refrigerators in another household through a direct transfer (giving it away or selling it) or indirect transfer (resale on the used appliance market).

Commonly implemented by third-party contractors (who collect and decommission participating appliances), these programs generate energy savings through the retirement of inefficient appliances. The decommissioning process captures environmentally harmful refrigerants and foam, and enables recycling of the plastic, metal, and wiring components.

[1] Secondary refrigerators are units not located in the kitchen.

2 Application Conditions of Protocol

Recycling programs currently have a range of designs:

- Recycling both primary and secondary refrigerators;
- Accepting only secondary refrigerators;
- Imposing restrictions on vintage eligibility;
- Offered in conjunction with point-of-sale rebates to encourage the purchase of ENERGY STAR®-rated refrigerators; and/or
- Offered as part of low-income, direct-install programs that install high-efficiency replacement units.[2]

The evaluation protocols described in this document, which pertain to all program variations listed, cover the energy savings from retiring operable-but-inefficient refrigerators. This protocol does not discuss the potential energy savings associated with the subsequent installation of a high-efficiency replacement refrigerator (which may occur as part of a separate retail products program).[3]

[2] Low-income, direct-install programs target refrigerators that otherwise would have continued to operate and replace them with comparably sized, new high-efficiency models. Therefore, the basis for estimating savings from these types of programs is different from the other program variations noted. This difference is discussed further in Section 3 of this chapter.

[3] As discussed under the section Considering Resource Constraints of the "Introduction" chapter to this UMP report, small utilities (as defined under the U.S. Small Business Administration (SBA) regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

3 Savings Calculations

This protocol provides guidance for estimating both gross and net savings.

The total gross energy savings[4] (kilowatt hours [kWhs] per year) achieved from recycling less efficient-but-operable refrigerators is calculated using the following general algorithm:

Equation 1

$$GROSS\_kWh = N * EXISTING\_UEC * PART\_USE$$

Where:

$GROSS\_kWh$	=	Annual electricity savings measured in kilowatt hours
N	=	The number of refrigerators recycled through the program
$EXISTING\_UEC$	=	The average annual unit energy consumption of participating refrigerators
$PART\_USE$	=	The portion of the year the average refrigerator would likely have operated if not recycled through the program

Due to the considerable potential for free-ridership in appliance recycling programs in general, this protocol includes a discussion of net savings. For this protocol, the net adjustment accounts for current early replacement and recycling practice. The total net energy savings (kWhrs/year) is calculated as follows:

Equation 2

$$NET\_kWh = N * NET\_FR\_SMI\_kWh$$

Where:

$NET\_FR\_SMI\_kWh$	=	Average per-unit energy savings net of naturally occurring removal from grid and secondary market impacts

The recommended techniques for estimating each of these parameters are described in the sections below.

[4] The evaluation protocol methods focus on energy savings; they do not include other parameter assessments such as peak coincidence factor (demand reduction), incremental cost, or measure life.

4 Measurement and Verification Plan

This section provides instructions for determining the parameters required to estimate a refrigerator recycling program's total gross savings (*GROSS_kWh*).

The key parameters are:

- Measure verification (*N*)
- Annual energy consumption (*EXISTING_UEC*)
- Part-use factor (*PART_USE*).

4.1 International Performance Measurement and Verification Protocol Option

Option B (Retrofit Isolation) is the recommended International Performance Measurement and Verification Protocol Option for this measure. Option B, which relies on short-term or continuous metering, most commonly includes both pre- and post-retrofit metering, or, when pre- retrofit metering is not possible, just post-metering. However, in the case of refrigerator recycling, evaluators should conduct just pre-retrofit metering to determine the energy consumption of the appliance that will be recycled by the program.

4.2 Measure Verification (*N*)

The program administrator or the third-party implementation contractor should record the number of refrigerators recycled through a program. Ideally, the data for all participating refrigerators are compiled electronically in a database that tracks the following information (at a minimum):

- Age (in years, or year of manufacture)
- Size (in cubic feet)
- Configuration (top freezer, bottom freezer, side-by-side, or single door)
- Date the refrigerator was removed
- Complete customer contact information.

This protocol recommends that early in the evaluation process, the evaluators review the program databases to ensure they are being fully populated and contain sufficient information to inform subsequent evaluation activities.

Self-reported verification of program recycling records via a survey of randomly sampled participants has proven to be a reliable methodology. Survey efforts should include a sufficient sample of participants to meet the required level of statistical significance. When no requirements exist, this protocol recommends a sample that achieves, at minimum, a 90% level of confidence with a ±10% margin of error. Past evaluations have shown that participants typically have little difficulty confirming the number of units recycled and the approximate date the removal took place (Cadmus 2010).

4.3 Annual Energy Consumption (*EXISTING_UEC*)

To determine the average per-unit annual energy consumption, use a regression-based analysis that relies on either:

- Metering a sample of participating units or
- Using metered data that was collected as part of other recycling program evaluations (when evaluation resources do not support primary data collection).

Average savings, as determined through either of these approaches, may be used but need to be updated at least every three years to account for program maturation.

This protocol strongly recommends that evaluators conduct a metering study, if possible. As this method is the preferred evaluation approach, the remainder of this section outlines the best practices for (1) implementing a metering study and (2) using the results to estimate annual energy consumption and, subsequently, energy savings.

4.3.1 About In Situ Metering

Historically, recycling evaluations have primarily relied on unit energy consumption (UEC) estimates from the U.S. Department of Energy (DOE) testing protocols (DOE 2008).[5] However, recent evaluations indicate that DOE test conditions (for example, empty refrigeration and freezers cabinets, no door openings, and 90˚F test chamber) may not accurately reflect UECs for recycled appliances (ADM 2008, Cadmus 2010). As a result, evaluations have increasingly used *in situ* (meaning "in its original place") metering to assess energy consumption.

In situ metering is recommended for two reasons:

- It accounts for environmental conditions and usage patterns within participating homes (for example, door openings, unit location, and exposure to weather), which are not explicitly accounted for in DOE testing.
- Most of the DOE-based UECs that are publicly available in industry databases were made at the time the appliance was manufactured, rather than when the unit was retired. Using testing data from the time of manufacture requires that assumptions be made about the degree of an appliance's degradation. *In situ* metering is conducted immediately prior to program participation (that is, at the time of the unit's retirement), so it is unnecessary to make such an adjustment or assumption.

In summary, while the DOE testing protocols provide accurate insights into the relative efficiency of appliances (most commonly at their time of manufacture), *in situ* metering yields the most accurate estimate of energy consumption (and, therefore, savings) for operable appliances that are less efficient.

[5] Evaluations have also used forms of billing analysis; however, this protocol does not recommend billing analysis or any other whole-house approach. The magnitude of expected savings—given total household energy consumption and changes in consumption unrelated to the program—could result in a less certain estimate than can be obtained from an end-use specific approach.

4.3.1.1 Key Factors for In Situ Metering

The following factors should be considered when implementing an *in situ* metering study:

- ***Sample Size.*** The recommended levels of statistical significance, which dictate the necessary sample size, are outlined in UMP *Chapter 11: Sample Design Cross-Cutting Protocol*. It is recommended that evaluators assume a minimum coefficient of variation of 0.5 to ensure that a sufficient sample is available to compensate for attrition issues that routinely occur in field measurement.[6] For refrigerators, these attrition issues may include simple meter failure, relocation of the unit during metering, and atypical usage (for example, the refrigerator is prematurely emptied in preparation for program pickup). This protocol recommends that evaluators educate study participants (and provide written leave-behind materials) about not relocating the refrigerator or otherwise using the unit in any manner inconsistent with historical usage.

- ***Stratification.*** The program theory assumes that most recycled appliances would have been used as secondary units had they not been decommissioned through the program.[7] However, some units may continue to operate as a primary unit within the same home. To correctly account for differences in usage patterns, it is critical to stratify the metering sample to represent the different usage types.[8]

  For programs evaluated previously, information may be available about the proportion of refrigerators likely to have been used as primary versus secondary units. If so, that information can be leveraged to develop stratification quotas for the metering study.

  Once established, strict quotas should be enforced during the recruitment process because participants who recycle secondary appliances are typically more willing to participate in a metering study than those who recycle primary appliances. Participants who are recycling their primary appliance are typically replacing them, and they are often unwilling to deal with the logistics related to rescheduling the delivery of their new unit.

  Additional stratification is not critical, due to the high degree of collinearity between refrigerator age, size, and configuration. However, when sufficient evaluation resources are available, targeting a sample of appliances with less common characteristics can reduce collinearity and increase the final model's explanatory power.

- ***Duration.*** To capture a range of appliance usage patterns, meters need to be installed for a minimum of 10 to 14 days.[9] Collecting approximately two weeks' worth of energy-consumption data ensures that the metering period covers weekdays and weekends. Longer metering periods will provide a greater range of usage (and more data points);

---

[6] For a broader discussion of the coefficient of variation, see the UMP Chapter 11.

[7] This includes several scenarios: The refrigerator may continue as a secondary appliance within the same home, be transitioned from a primary to a secondary appliance within the same home, or become a secondary unit in another home.

[8] This protocol recommends stratification by usage type, even for programs that only accept secondary units, as primary units are typically still recycled through these programs (via gaming or confusion about requirements).

[9] The previously cited evaluations in California (ADM 2008, Cadmus 2010) both collected metering data for a minimum of from 10 to 14 days.

however, the duration needs to be balanced with the customers' desire to have their refrigerator removed and recycled.

- ***Equipment.*** To capture information on compressor cycling, record the data in intervals of five minutes or less. If the meters' data capacity permits, a shorter interval (of one or two minutes) is preferable. When possible, meter the following parameters; however, if metering efforts are limited, prioritize the parameters in this order:
  - Current and/or power
  - Internal refrigerator and/or freezer cabinet temperature
  - Ambient temperature
  - Frequency and duration of door openings.[10]

  Not all the aforementioned metered values are used to determine energy consumption. Some help identify potential problems in the metering process, and thus increase the quality of the data. (For example, a comparison of ambient room temperature to internal cabinet temperature can be used to determine if the appliance was operational throughout the entire metering period.) This protocol recommends that evaluators perform similar diagnostics on all raw metering data before including an appliance in the final analysis dataset.

- ***Seasonality.*** Previous metering studies have shown that the energy consumption of secondary appliances in unconditioned spaces differs by season—especially in regions that experience extreme summer and/or winter weather.[11] As a result, metering needs to be conducted in waves on separate samples. By capturing a range of weather conditions using multiple metering waves (which include winter and summer peaks, as well as shoulder seasons), it is possible to annualize metering results more accurately. If it is not possible to meter appliances during multiple seasons, then annualize the metered data using existing refrigerator load shapes (that are utility specific, when available) to avoid producing seasonally biased estimates of annual unit consumption.

- ***Recruitment.*** When arranging for metering, evaluators must contact participating customers before the appliance is removed. By working closely with the program implementers (who can provide daily lists of recently scheduled pickups), evaluators can contact those customers to determine their eligibility and solicit their participation in the metering study.

  This protocol recommends providing incentives to participants. Incentives aid in recruitment because they both provide recognition of the participants' cooperation and offset the added expense of continuing to operate their refrigerator during metering.

---

[10] The Cadmus 2010 evaluation used the following metering equipment: HOBO U9-002 Light Sensor (recorded the frequency and duration of door openings), HOBO U12-012 External Data Logger (recorded the ambient temperature and humidity), HOBO U12-012 Internal Data Logger (recorded the cabinet temperature), HOBO CTV-A (recorded the current), and the Watts up? Pro ES Power Meter (recorded energy consumption).

[11] *Michigan Energy Efficiency Measure Database* memo by Cadmus regarding Consumers Energy and DTE Energy appliance recycling programs.

Once participants are recruited, the evaluator and the implementer should collaborate in scheduling the participants' pickup after all the metering equipment is removed.

- ***Installation and Removal.*** Evaluators can install and remove all metering equipment, or, to minimize costs, program implementers can perform these functions. However, when program implementers are involved in the metering process, the evaluator must still independently conduct all sampling design and selection, recruitment, metering equipment programming, data extraction, and data analysis.

  To ensure installations and removals are performed correctly, evaluators should train the implementers' field staff members and, ideally, accompany them on a sample of sites. If time and evaluation resources permit, evaluators should verify early in the data collection phase that the metering equipment is installed properly at a small sample of participating homes, in order to identify and correct any installation issues.

  Because the metering process requires an additional trip to customer homes, evaluators need to compensate the implementers for their time. Consequently, the evaluators should contact implementers as early as possible to determine the viability of this approach and agree upon the appropriate compensation.

- ***Frequency.*** Because the characteristics of recycled refrigerators change as a program matures and greater market penetration is achieved, metering should be conducted approximately every three years. Savings estimates that rely exclusively on metering data older than three years reflect the current program year inaccurately. This is most commonly due to changes in the mix of recycled appliances manufactured before and after the establishment of appliance-related standards (including various state, regional, or federal standards) between program years. The main impact of these changes is a long-term downward effect on the savings associated with recycling programs.

4.3.2 About Regression Modeling

To estimate the annual UEC of the average recycled refrigerator, this protocol recommends that evaluators use a multivariate regression model that relates observed energy consumption to refrigerator characteristics.

Evaluators should employ models that use daily or hourly observed energy consumption as the dependent variable. Independent variables should include key refrigerator characteristics or environmental factors determined to be statistically significant. This functional form allows the coefficient of each independent variable to indicate the relative influence of that variable (or appliance characteristic) on the observed energy consumption, holding all other variables constant. This approach allows evaluators to estimate the energy consumption of all participating appliances based on the set of characteristics maintained in the program tracking database.

In estimating UEC, both time and cross-sectional effects must be accounted for. This can be done in one of two ways:

- ***Use a model that simultaneously estimates the impacts of longitudinal (time) and cross-sectional effects on energy consumption.*** This approach is recommended if the sample size is reasonably large and if units are observed across both summer and winter peak periods.

- ***Use a set of time-series models.*** If metering is done during only the winter or summer, use a refrigerator load shape from a secondary source to extrapolate the annual UEC for each metered refrigerator. Then apply a regression model using the entire metering sample to predict annualized consumption as a function of cross-sectional variables.

Once model parameters are estimated, the results may be used to estimate UEC for each refrigerator recycled through a program, based on each unit's unique set of characteristics. An example is provided later in this section.

The exact model specification (a set of appliance characteristics or independent variables) yielding the greatest explanatory power varies from study to study, based on the underlying metering data. Thus, this protocol does not mandate a certain specification. However, evaluators should consider—at a minimum—the following independent variables:

- Age (years) and corresponding vintage (compliance with relevant efficiency code)
- Size (in cubic feet)
- Configuration (top freezer, bottom freezer, side-by-side, or single door)
- Primary or secondary designation
- Conditioned or unconditioned space[12]
- Location (kitchen, garage, basement, porch, etc.)
- Weather (cooling degree days [CDD] and heating degree days [HDD]).

For each set of potential independent variables, evaluators should assess the variance inflation factors, adjusted R^2, residual plots, and other measures of statistical significance and fit.

In the specification process, evaluators should also consider the following elements:

- Estimating model parameters by using an ordinary least squares or generalized least squares method
- Transforming explanatory variables (logged and squared values) based on theoretical and empirical methods
- Considering interaction terms (such as between refrigerators located in unconditioned spaces and CDD/HDD) when they are theoretically sound (that is, not simply to increase the adjusted R^2 or any other diagnostic metric)
- Balancing model parsimony with explanatory power (It is very important not to over-specify the model(s). As the regression models are used to predict consumption for a wide variety of units, overly specified models can lose their predictive validity.).

[12] The primary or secondary designation and conditioned or unconditioned space variables may exhibit a strong collinearity; consequently, do not include both in the final model.

The following example regression model is based on data from 472 refrigerators metered and recycled through five utilities:[13]

Table 1. Example Refrigerator UEC Calculation Using Regression Model and Program Values

Independent Variable	Estimated Coefficient (Daily kWh)	Program Values (Average/Proportion)
Intercept	0.582	-
Appliance Age (years)	0.027	22.69
Dummy: Manufactured Pre-1990	1.055	0.63
Appliance Size (square feet)	0.067	18.92
Dummy: Single-Door Configuration	-1.977	0.06
Dummy: Side-by-Side Configuration	1.071	0.25
Dummy: Primary Usage Type (in absence of program)	0.6054	0.36
Interaction: Located in Unconditioned Space * CDDs	0.020	2.49
Interaction: Located in Unconditioned Space * HDDs	-0.045	1.47
Estimated UEC (kWh/Year)		1,240

Once the characteristics of a specific appliance are determined, they should be substituted in the equation to estimate the UEC for that appliance. After the UEC is calculated for each participating unit, a program average UEC can be determined.

$$\begin{aligned}Refrigerator\ UEC &= 365.25 \\ &* [0.582 + 0.027 * (22.69\ years) + 1.055 \\ &* (63\%\ manufactured\ before\ 1990) + 0.067 * (18.92\ cubic\ feet) \\ &- 1.977 * (6\%\ single\ door\ units) + 1.071 * (25\%\ side - by - side) \\ &+ 0.605 * (36\%\ primary\ usage) + 0.02 \\ &* (2.49\ unconditioned\ CDDs) - 0.045 * (1.47\ unconditioned\ HDDs)] \\ &= 1{,}240\ annual\ kWh\end{aligned}$$

Similarly, the following sample regression is based on 57 freezers metered through three utilities.

[13] The example data are based on metering from the previously cited evaluations Cadmus 2010 and *Michigan Energy Efficiency Measure Database* memo collected between 2008 and 2012.

Table 2. Example Freezer UEC Calculation Using Regression Model and Program Values

Independent Variable	Estimate Coefficient (Daily kWh)	Program Values (Average/Proportion)
Intercept	-0.955	-
Appliance Age (years)	0.045	30.94
Dummy: Manufactured Pre-1990	0.543	0.71
Appliance Size (square feet)	0.12	17.27
Dummy: Chest Configuration	0.298	0.19
Interaction: Located in Unconditioned Space * HDDs	-0.031	11.18
Interaction: Located in Unconditioned Space * CDDs	0.082	1.84
Estimated UEC (kWh/Year)		1,007

Substituting the program averages into the equation we get:

$$\begin{aligned} Freezer\ UEC &= 365.25 \\ &* [-0.955 + 0.045 * (30.94\ years) + 0.543 \\ &* (71\%\ manufactured\ before\ 1990) + 0.12 * (17.27\ cubic\ feet) \\ &+ 0.298 * (19\%\ chest\ freezer\ units) - 0.031 \\ &* (11.18\ unconditioned\ HDDs) + 0.082 \\ &* (1.84\ unconditioned\ CDDs)] = 1{,}007\ annual\ kWh \end{aligned}$$

4.3.3 Using Secondary Data

When evaluation resources do not support *in situ* metering, evaluators should leverage a model developed through the most appropriate *in situ* metering-based evaluation undertaken for another utility. The most appropriate study will be comparable to the program being evaluated in terms of the following factors:

- Age of the study (recent is most desirable)
- Similar average appliance characteristics (comparable sizes, configurations, etc.)
- Similar geographical location (due to differences in climate)
- Similar customer demographics (due to differences in usage patterns).

Use the aggregated UEC model presented in Table 1 or Table 2 when (1) *in situ* metering is not an option *and* (2) a recently developed model from a single comparable program cannot be identified. It is important to note, however, that the meter data used to develop the example regression models above was collected between 2009 and 2012. As time passes, those metered units become less representative of appliances being recycled through current programs.

4.4 Part-Use Factor (*PART_USE*)

"Part-use" is an appliance recycling-specific adjustment factor used to convert the UEC (determined through the methods detailed above) into an average per-unit gross savings value. The UEC itself is not equal to the gross savings value for two reasons:

- The UEC model yields an estimate of annual consumption
- Not all recycled refrigerators would have operated year-round had they not been decommissioned through the program.

Table 3 provides a summary of the three part-use categories, each with its own part-use factor. The part-use factors for refrigerators that would have run full-time (1.0) and those that would have not run at all (0.0) are consistent across evaluations. The part-use factor for refrigerators that would have been used for a portion of the year varies by program (and is between 0.0 and 1.0). For example, a refrigerator estimated to operate a total of three months over the course of a year (most commonly to provide additional storage capacity during the holidays) would have a part-use factor of 0.25.

Table 3. Part-Use Factors by Category

Part-Use Category	Part-Use Factor
Likely to not operate at all in absence of the program	0
Likely to operate part of the time in absence of the program	0 to 1
Likely to operate year-round in absence of the program	1

Using participant surveys, evaluators should determine the number of recycled units in each part-use category, as well as the portion of the year that the refrigerators that *would have been used part of the time* were likely to have been operated. The protocol recommends handling this assessment through the following multistep process:

1. ***Ask participants where the refrigerator was located for most of the year prior to being recycled.*** By asking about the refrigerator's long-term location, evaluators can obtain more reliable information about the unit's usage and can avoid using terms that often confuse participants (such as primary and secondary), especially when replacement occurs. It is recommended that evaluators designate all refrigerators previously located in a kitchen as primary units and those previously in any other location as secondary units.

   Note that it is important not to ask about the refrigerator's location when it was collected by the program implementer, as many units are relocated to accommodate the arrival of a replacement appliance or to facilitate program pickup.

2. ***Ask those participants who indicated recycling a secondary refrigerator whether the refrigerator was unplugged, operated year-round, or operated for a portion of the preceding year.*** Evaluators can assume that all primary units were operated year-round.

3. ***Ask those participants who indicated that their secondary refrigerator was operated for only a portion of the preceding year to estimate the total number of months during that year the refrigerator was plugged in.*** Then divide the average number of months

specified by this subset of participants by 12 to calculate the part-use factor for all refrigerators operated for only a portion of the year.

Three steps enable evaluators to obtain important and specific information about how a refrigerator was used before it was recycled. The example program details provided in Table 4 show that:

- The participant survey determined that 93% of recycled refrigerators were operated year-round either as a primary or secondary unit. (Again, the part-use factor associated with these refrigerators is 1.0.)
- Four percent of refrigerators were not used at all in the year before being recycled. The part-use factor associated with this portion of the program population is 0.0, and no energy savings are generated by the refrigerator's removal and eventual decommissioning.
- The remaining refrigerators (3%) were operational for a portion of the year. Specifically, the survey determined that part-time refrigerators were operated for an average of three months per year (indicating a part-use factor of 0.25).

Using this information, evaluators should calculate the overall part-use factors for secondary units only, as well as for all recycled units. These factors are derived by applying a weighted average of the adjusted part-use per-unit energy savings for each part-use category. This calculation uses the UEC determined through the methods described in the "About Regression Modeling" section. In this example, the program's secondary-only part-use factor is 0.88, while the overall part-use factor is 0.93. Again, the part-use values provided in the table below are illustrative only; evaluators should use the process described above to calculate program-specific part-use factors as part of all evaluations.

Table 4. Example Calculation of Historical Part-Use Factors

Usage Type and Part-Use Category	Percentage of Recycled Units	Part-Use Factor	Per-Unit Energy Savings (kWh/Yr)
Secondary Units Only			
Not in Use	6%	0.00	-
Used Part-Time	8%	0.25	310
Used Full-Time	86%	1.00	1,240
Weighted Average	**100%**	**0.88**	**1,091**
All Units (Primary and Secondary)			
Not in Use	4%	0.00	-
Used Part-Time	3%	0.25	310
Used Full-Time	93%	1.00	1,240
Weighted Average	**100%**	**0.93**	**1,163**

Next, evaluators should combine these historically observed part-use factors with participants' self-reported action had the program *not* been available. (That is, the participants' report as to whether they would they have kept or discarded their refrigerator.)[14]

The example provided in Table 5 demonstrates how a program's part-use factor is determined using a weighted average of historically observed part-use factors and participants' likely action in the absence of the program.[15] Here, the result is a part-use value of 0.91, based on the expected future use of the refrigerators had they not been recycled.

Table 5. Example Calculation of Prospective Program Part-Use

Use Prior to Recycling	Likely Use Independent of Recycling	Part-Use Factor	Percentage of Participants
Primary	**Kept (as primary unit)**	**1.0**	**15%**
	Kept (as secondary unit)	0.88	25%
	Discarded	0.93	15%
Secondary	Kept	0.88	30%
	Discarded	0.93	15%
Overall	**All**	**0.91**	**100%**

Applying the determined prospective part-use factor (*PART_USE*) of 0.91 to the determined annual energy consumption (*EXISTING_UEC*) of 1,240 kWh per year yields the program's average per-unit gross savings (which, in this case, is 1,128 kWh per year).

Recent evaluations of appliance recycling programs have determined that part-use factors typically range from 0.85 to 0.95 (Navigant 2010). Newer appliance recycling programs have exhibited a part-use factor at the lower end of this range. This is attributed to that fact that many unused or partially used appliances sat idle before the program launch simply because participants lacked the means to discard them. (The recycling program then provided the means.) In addition, the newer programs tend to focus on collecting secondary units (which are subject to partial use), while mature programs tend to focus on avoided retention (replacing primary appliances). As a result, part-use factors tend to increase over time.

The part-use factor should be reassessed annually for newer programs because it may change more rapidly during the early stages of a program's lifecycle. After a program has been in operation for at least three years, it is sufficient to conduct a part-use assessment every other year.

[14] Since the future usage type of discarded refrigerators is unknown, evaluators should apply the weighted part-use average of all units (0.93) to refrigerators that would have been discarded independent of the program. This approach acknowledges that discarded appliances might be used as primary or secondary units in the would-be recipient's home.

[15] Evaluators should not calculate part-use using participant's estimates of future use had the program not been available. Historical estimates based on actual usage rates are more accurate, especially because it is possible that some participants will underestimate future usage (believing they will only operate the unit for part of the year, despite the fact that most refrigerators are operated continuously once plugged in).

4.5 Refrigerator Replacement

In most cases, the per-unit gross energy savings attributable to the program is equal to the energy consumption of the recycled appliance (rather than being equal to the difference between the consumption of the participating appliance and its replacement, when applicable). This is because the energy savings generated by the program are not limited to the change within the participant's home but rather to the total change in energy consumption at the grid level.

This concept is best explained with an example. Suppose a customer decides to purchase a new refrigerator to replace an existing one. When the customer mentions this to a neighbor, the neighbor asks for that existing refrigerator to use as a secondary unit. The customer agrees to give the existing appliance to the neighbor; however, before this transfer is made, the customer learns about a utility-sponsored appliance recycling program. The customer decides to participate in the program because the incentive helps offsets the cost of the new refrigerator. As a result of program intervention, the customer's appliance is permanently removed from operation in the utility's service territory.

From the utility's perspective, the difference in grid-level energy consumption—and the corresponding increase in program savings—are equal to the consumption of the recycled appliance *and not* to the difference between the energy consumption of the participating appliance and its replacement. In this example, it is important to note that the participant planned to replace the appliance.

In general, the purchase of new refrigerators is part of the naturally occurring appliance lifecycle, typically independent of the program and tantamount to refrigerator load growth. It is not the purpose of the program to prevent these inevitable purchases, but rather to minimize the grid-level refrigerator load growth by limiting the number of existing appliances that continue to be operated once they are replaced.

It may be possible that a recycling program could induce a replacement (that is, the participant would ***not*** have purchased the new refrigerator in absence of the recycling program). However, accurately measuring this effect is difficult, as surveyed program participants are often confused by the questions about their intentions to purchase another appliance absent the recycling program. Past evaluations that assessed induced replacement rates found that customers may initially state that they would not have purchased a replacement absent the influence of a program. But when asked follow-up questions about which aspect of the program influenced their decision, customers often mentioned factors unrelated to the program, such as wanting a larger appliance or that the prior unit was old and they wanted something more dependable. These follow-up answers contradict the earlier answers that the program was the reason they decided to acquire a new unit. Given the relatively small impact of potential program induced replacements, the difficulty in measuring accurately, and the general confusion resulting from these questions during customer surveys, this protocol recommends against estimating induced replacement and using as a factor when determining net program savings.

Appliances that, independent of the program, would have been discarded in a way leading to destruction (such as being taken to a landfill)—rather than being transferred to a new user—are captured by the program net-to-gross (NTG) ratio. Thus, no net savings are generated by the program. This is a separate issue from estimating gross energy savings and is discussed in Section 5 in more detail.

5 Net Savings

This section provides instructions for determining the additional parameters required to estimate a refrigerator recycling program's net savings (*NET_kWh*). In the case of refrigerator recycling, net savings are only generated when the recycled appliance would have continued to operate absent program intervention (either within the participating customer's home or at the home of another utility customer).

5.1 Free-Ridership and Secondary Market Impacts (*NET_FR_SMI_kWh*)

To estimate free-ridership and secondary market impacts, this protocol recommends using a combination of the responses of surveyed participants, surveyed nonparticipants, and (if possible) secondary market research. Use all these data together to populate a decision tree of all possible savings scenarios. Then take a weighted average of these scenarios to calculate the savings that can be credited to the program after accounting for either free-ridership or the program's interaction with the secondary market. Populate this decision tree based on what the participating households would have done outside the program and, if the unit would have been transferred to another household, whether the would-be acquirer of that refrigerator finds an alternate unit instead.

In general, independent of program intervention, participating refrigerators would have been subject to one of the following scenarios:

1. The refrigerator would have been kept by the household
2. The refrigerator would have been discarded by a method that transfers it to another customer for continued use
3. The refrigerator would have been discarded by a method leading to its removal from service.

These scenarios encompass what has often been referred to as free-ridership (the proportion of units that would have been taken off the grid absent the program).

For units that would have been transferred to another household, the question then becomes what purchasing decisions are made by the would-be acquirers of participating units now that these units are unavailable:

1. They could not purchase or acquire another unit
2. They could purchase or acquire another unit.

Adjustments to savings based on these factors are referred to as the program's secondary market impacts.

5.1.1 Free-Ridership

The first step is to estimate the distribution of participating units likely to have been kept or discarded absent the program. Further, there are two possible scenarios for discarded units, so in total, there are three possible scenarios independent of program intervention:

1. Unit is discarded and transferred to another household
2. Unit is discarded and destroyed
3. Unit is kept in the home.

As participants often do not have full knowledge of the available options for and potential barriers to disposing refrigerators (Scenarios 1 and 2), this document recommends using nonparticipant[16] survey data to mitigate potential self-reporting errors. The proportion of units that would have been kept in the home (Scenario 3) can be estimated exclusively through the participant surveys, as participants can reliably provide this information.

Nonparticipant surveys provide information from other utility customers regarding how they actually discarded their refrigerator independent of the program. Evaluators can use this information to estimate the proportion of discarded units that are transferred (Scenario 1) versus destroyed (Scenario 2).

Specifically, evaluators should calculate the distribution of the ratio of likely discard scenarios as a weighted average from both participants and nonparticipants (when nonparticipant surveys are possible). The averaging of participant and nonparticipant values mitigates potential biases in the responses of each group.[17] As the true population of nonparticipants is unknown, the distribution should be weighted using the inverse of the variance of participant and nonparticipant free-ridership ratios.[18] This method of weighting gives greater weight to values that are more precise or less variable. As demonstrated in Table 6,[19] this approach results in an estimate of the proportion of participating appliances that would have been permanently destroyed (Scenario 1), transferred to another user (Scenario 2), or kept (Scenario 3).

Table 6. Determination of Discard and Keep Distribution

Discard /Keep	Proportion of Participant Sample	Sample	Discard Scenario	N	SE	Weight	Proportion of Discards	Overall Proportion
Discard	70%	Participant	Transfer	70	0.05	0.60	80%	
			Destroy				20%	
		Nonparticipant	Transfer	70	0.06	0.40	60%	
			Destroy				40%	
		Weighted Average	Transfer				72%	50%
			Destroy				28%	20%
Kept	30%							30%

16 "Nonparticipants" are defined as utility customers who disposed of an operable refrigerator outside of the utility program while the program was being offered.
17 Participant responses may be biased due to not fully understanding barriers to various disposal options. Nonparticipant decisions may not be representative of what participants would do in the absence of the program due to participants self-selecting into the program (as opposed to being randomly enrolled).
18 Inverse variance weights involve weighting each estimate by the inverse of its squared standard error. This technique is common in meta-analysis literature and is used to place greater weight on more reliable estimates.
19 More detail on how this information is used to determine net savings can be found in "Section 6, Summary Diagram."

Another option for informing free-ridership is to interview local used appliance dealers. These interviews can help determine the viability of selling used refrigerators of given ages or set of characteristics. This information can provide insight into the transferability of these units and supplement the self-reported actions of participants had they not participated in the recycling program.

5.1.1.1 Participant Self-Reported Actions

To determine the percentage of participants in each of the three scenarios, evaluators should begin by asking surveyed participants about the likely fate of their recycled appliance had it not been decommissioned through the utility program. Responses provided by participants can be categorized as follows:

- Kept the refrigerator
- Sold the refrigerator to a private party (either an acquaintance or through a posted advertisement)
- Sold or gave the refrigerator to a used appliance dealer
- Gave the refrigerator to a private party, such as a friend or neighbor
- Gave the refrigerator to a charity organization, such as Goodwill Industries or a church
- Had the refrigerator removed by the dealer from whom the new or replacement refrigerator was obtained
- Hauled the refrigerator to a landfill or recycling center
- Hired someone else to haul the refrigerator away for junking, dumping, or recycling.

To ensure the most reliable responses possible and to mitigate socially desirable response bias, evaluators should ask some respondents additional questions. For example, participants may say they would have sold their unit to a used appliance dealer. However, if the evaluation market research revealed that used appliance dealers were unlikely to purchase the older unit (due to its age or condition), then participants should be asked what they would have likely done *had they been unable to sell the unit to a dealer*. Evaluators should then use the response to this question in assessing free-ridership.

If market research reveals that local waste transfer stations charge a fee for dropping off refrigerators, inform participants about the fee if they initially specify this as their option and then ask them to confirm what they would have done in the absence of the program. Again, evaluators should use this response to assess free-ridership.

Use this iterative approach with great care. It is critical that evaluators find the appropriate balance between increasing the plausibility of participants' stated actions (by offering context that might have impacted their decision) while not upsetting participants by appearing to invalidate their initial response.

Next, evaluators should assess whether each participant's final response indicates free-ridership.

- Some final responses clearly indicate free-ridership, such as: "I would have taken it to the landfill or recycling center myself."
- Other responses clearly indicate no free-ridership, such as when the refrigerator would have remained active within the participating home ("I would have kept it and continued to use it") or used elsewhere within the utility's service territory ("I would have given it to a family member, neighbor, or friend to use.").

5.1.2 Secondary Market Impacts

If it is determined that the participant would have directly or indirectly (through a market actor) transferred the unit to another customer on the grid, the next question addresses what that potential acquirer did when that unit was unavailable. There are three possibilities:

A. ***None of the would-be acquirers would find another unit.*** That is, program participation would result in a one-for-one reduction in the total number of refrigerators operating on the grid. In this case, the total energy consumption of avoided transfers (participating appliances that otherwise would have been used by another customer) should be credited as savings to the program. This position is consistent with the theory that participating appliances are essentially convenience goods for would-be acquirers. (That is, the potential acquirer would have accepted the refrigerator had it been readily available, but would not seek out an alternate unit because the refrigerator was not a necessity.)

B. ***All the would-be acquirers would find another unit.*** Thus, program participation has no effect on the total number of refrigerators operating on the grid. This position is consistent with the notion that participating appliances are necessities and that customers will always seek alternative units when participating appliances are unavailable.

C. ***Some of the would-be acquirers would find another unit, while others would not.*** This possibility reflects the awareness that some acquirers were in the market for a refrigerator and would acquire another unit, while others were not (and would only have taken the unit opportunistically).

It is difficult to answer this question with certainty, absent utility-specific information regarding the change in the total number of refrigerators (overall and used appliances specifically) that were active before and after program implementation. In some cases, evaluators have conducted in-depth market research to estimate both the program's impact on the secondary market *and* the appropriate attribution of savings for this scenario. Although these studies are imperfect, they can provide utility-specific information related to a program's net energy impact. Where feasible, evaluators and utilities should design and implement such an approach. Unfortunately, this type of research tends to be cost-prohibitive, or the necessary data may simply be unavailable.

Because the data to inform such a top-down market-based approach may be unavailable, evaluators have employed a bottom-up approach that centers on identifying and surveying recent acquirers of non-program used appliances and asking what they would have done had the specific used appliance they acquired not been available. While this approach results in quantitative data to support evaluation efforts, it is uncertain if:

- The used appliances these customers acquired are in fact comparable in age and condition to those recycled through the program

- These customers can reliably respond to the hypothetical question.

Further, any sample composed entirely of customers who recently acquired a used appliance seems inherently likely to produce a result that aligns with the second possibility (B) presented above.

As a result of these difficulties and budget limitations, this protocol recommends the last possibility (C) when primary research cannot be undertaken. Specifically, evaluators should assume that half (0.5, the midpoint of possibilities A and B) of the would-be acquirers of avoided transfers found an alternate unit.

Once the proportion of would-be acquirers who are assumed to find an alternate unit has been determined, the next question is whether the alternate unit was likely to be another used appliance (similar to those recycled through the program) or, with fewer used appliances presumably available in the market due to program activity, would the customer acquire a new standard-efficiency unit instead.[20] For the reasons previously discussed, it is difficult to estimate this distribution definitively. Thus, this protocol recommends a midpoint approach when primary research is unavailable: evaluators should assume that half (0.5) of the would-be acquirers of program units would find a similar, used appliance and half (0.5) would acquire a new, standard-efficiency unit.[21]

Figure 1 details the methodology for assessing the program's impact on the secondary market and applying the recommended midpoint assumptions when primary data are unavailable. As evident in the figure, accounting for market effects results in three savings scenarios: full savings (i.e., per-unit gross savings), no savings, and partial savings (i.e., the difference between the energy consumption of the program unit and the new, standard-efficiency appliance acquired instead).

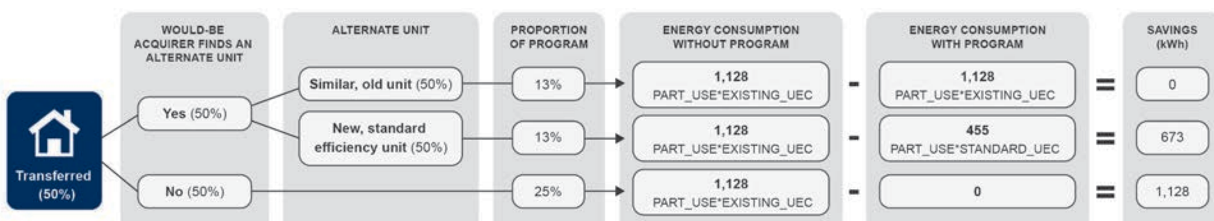


Figure 1. Secondary market impacts

[20] It is also possible the would-be acquirer of a program unit would select a new ENERGY STAR unit as an alternate. However, this protocol recommends that evaluators assume that any such used appliance supply restricted upgrades be limited to new, standard-efficiency units because (1) it seems most likely that a customer in the market for a used appliance would upgrade to the new lowest price point and (2) excluding ENERGY STAR units avoids potential double counting between programs when utilities offer concurrent retail rebates.

[21] Evaluators should determine the energy consumption of a new, standard-efficiency appliance using the ENERGY STAR website. Specifically, evaluators should average the reported energy consumption of new, standard-efficiency appliances of comparable size and similar configuration to the program units.

5.1.3 Integration of Free-Ridership and Secondary Market Impacts

Once the parameters of the free-ridership and secondary market impacts are estimated, a decision tree can be used to calculate the average per-unit program savings net of their combined effect. Figure 2 shows how these values are integrated into a combined estimate (*NET_FR_SMI_kWh*), shown on a per-unit basis.

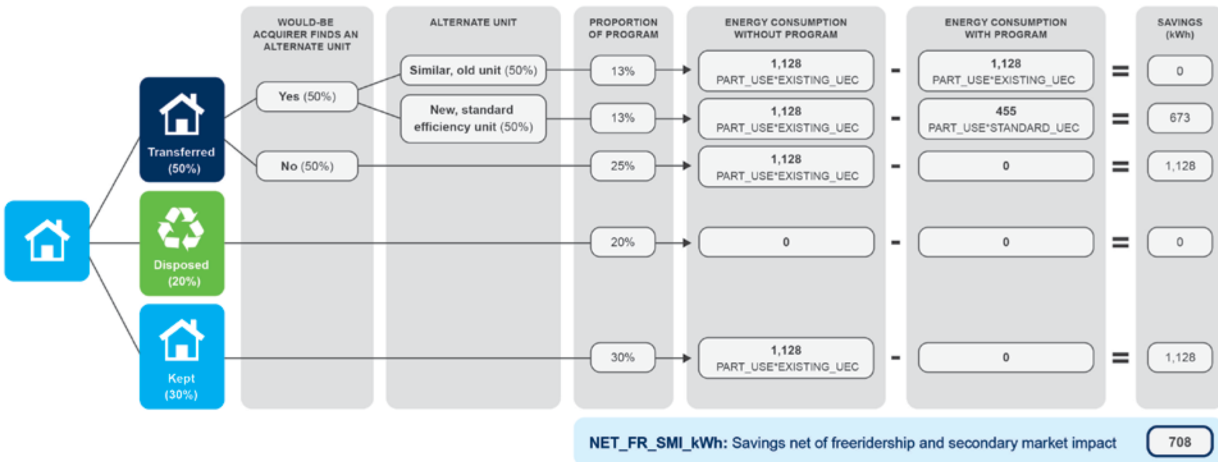


Figure 2. Savings net of free-ridership and secondary market impacts

As shown above, evaluators should estimate per-unit *NET_FR_SMI_kWh* by calculating the proportion of the total participating units associated with each possible combination of free-ridership and secondary market scenarios and its associated energy savings.

5.2 Spillover

This protocol does not recommend quantifying and applying participant spillover to adjust net savings for the following reasons:

- Unlike a CFL program, the opportunities for like spillover (the most common and defensible form of spillover for most downstream demand-side management programs) are limited in a recycling program because the number of refrigerators available for recycling in a typical home is limited.
- Unlike a whole-house audit program, recycling programs typically do not provide comprehensive energy education that would identify other efficiency opportunities within the home (and generate non-like spillover).
- Quantifying spillover accurately is challenging and, despite well-designed surveys, uncertainty often exists regarding the attribution of subsequent efficiency improvements to participation in the recycling program.

However, as a result of the ease of participation and high levels of participant satisfaction, appliance recycling programs may encourage utility customers to enroll in other available residential programs. While this is a positive attribute of recycling programs within a residential portfolio, all resulting savings are captured by other program evaluations.

5.3 Data Sources

After determining a program's gross energy savings, the net savings are determined by applying a NTG adjustment using several data sources:[22]

- ***Participant Surveys***. Surveys with a random sample of participants offer self-report estimates regarding whether participating refrigerators would have been kept or discarded independent of the program.[23] When participants indicate that the recycled refrigerator would have been discarded, ask for further details as to their likely method of disposal in the absence of the program. For example, ask whether the appliance would have been given to a neighbor, taken to recycling center, or sold to a used appliance dealer.
- ***Nonparticipant Surveys***. To mitigate potential response bias,[24] this protocol recommends using nonparticipant surveys to obtain information for estimating NTG. Information about how nonparticipants actually discarded their operable refrigerators outside of the program can reveal and mitigate potential response bias from participants. (Participants may overstate the frequency with which they would have recycled their old-but-operable refrigerator because they respond with what they perceive as being socially acceptable answers.) Nonparticipants, however, can only provide information about how units were actually discarded.[25] Because nonparticipant surveys require greater evaluation resources, it is acceptable to use smaller sample sizes.[26,27]
- ***Market Research***. Some participant and nonparticipant responses require additional information for determining definitively whether the old-but-operable refrigerator would have been kept in use absent the program. Responses requiring follow-up include:
    - "I would have sold it to a used appliance dealer"
    - "I would have had the dealer who delivered my new refrigerator take the old refrigerator."

  To inform a more robust NTG analysis, conduct market research by interviewing senior management from new appliance dealers and used appliance dealers (both local chains and big-box retailers). Ask about the viability of recycled refrigerators being resold on the used market had they not been decommissioned through the program. For example, do market actors resell none, some, or all picked-up refrigerators? If only some are resold, what are characteristics (for example, age, condition, features) that determine when a refrigerator is for resale. Information gained through this research (which should be conducted before the participant surveys) can be used to assess the reasonableness of

[22] When it is cost-prohibitive to survey nonparticipants and interview market actors, calculate free-ridership using participant surveys and secondary data from a comparable set of market actors.

[23] As noted previously, the number of participant surveys should be sufficient to meet the required level of statistical significance. A minimum of 90% confidence with ±10% precision is suggested.

[24] See UMP Chapter 11 for a broader discussion of sources of bias.

[25] Information regarding the likelihood that the recycled refrigerator would have been retained independent of program intervention can be obtained reliably through the participant surveys.

[26] The cost of identifying nonparticipants can be minimized by adding the nonparticipant NTG module to concurrent participant surveys for evaluations of other utility programs within the portfolio.

[27] For a general discussion of issues related to conducting surveys, see UMP Chapter 11.

participants' self-reported hypothetical actions independent of the program. This information can also be used to prompt participants to offer alternative hypothetical actions.[28]

[28] More detail is provided in the Section 5.1.1, "Free-Ridership."

6 Other Evaluation Issues

6.1 Remaining Useful Life

It is difficult to determine the number of years that a recycled refrigerator would have continued to operate absent the program and, therefore, the longevity of the savings generated by recycling old-but-operable refrigerators through the program. Participant self-reports are speculative and cannot account for unexpected appliance failure. Also, the standard evaluation measurements of remaining useful life (RUL) are not applicable, as most participating refrigerators are already past their effective useful life (EUL) estimates.

More primary research is needed on this topic to identify a best practice. In the interim and in lieu of a formal recommendation, this protocol offers two examples of estimation methods.

- RUL can be estimated as a function of a utility's new refrigerator EUL, using the following formula:[29] RUL = EUL/3
- RUL can be estimated using survival analysis (when appropriate data are available).[30]

6.2 Freezers

Although this protocol is focused on refrigerators, most utility appliance recycling programs also decommission stand-alone freezers. While differences exist between the evaluation approach for each appliance type (for example, all stand-alone freezers are secondary units, while refrigerators may be primary or secondary units), this protocol can also be used to evaluate the savings for freezers.

6.3 Customer Satisfaction

As recycling program mature, they experience a decrease in per-unit savings over time as the stock of available refrigerators to be recycled by the program were manufactured more recently and are therefore more efficient. The average age of appliances recycled tends to decline over time and fewer participating appliances were manufactured prior to the introduction of national appliance efficiency standards. This often leads to recycling programs being only marginally cost-effective.

While cost-effectiveness is an important criterion for considering whether to continue to offer a program, there are specific benefits generated by recycling programs beyond cost-effectiveness that are worth consideration.

Recycling programs tend to have very high levels of customer satisfaction. Evaluations have consistently found that well over 90% of customers are very satisfied with their experience. This gives utilities an opportunity to leave a positive impression with customers.

---

[29] This formula was obtained from the *Database for Energy Efficient Resources* (http://www.energy.ca.gov/deer/).

[30] In an evaluation of the NV Energy appliance recycling program, ADM Associates used survival analysis using secondary data from the 2009 California RASS. This involved estimating hazard rates for refrigerators based on the observed destruction of appliances at various ages. Once the hazard rate function was estimated, a table of expected RULs at each age was calculated. Where feasible, this approach should be followed using data specific to the given utility service area.

Additionally, recycling programs reach large numbers of customers who may be harder to reach via more traditional residential program offerings. Participants' positive experience with the recycling program provides an opportunity for the utility to educate them about energy efficiency and cross-promote other program offerings, which in turn can increase participation in other efficiency programs.

7 Looking Forward

As mentioned in Section 0, the metering data used to develop the regression models in this protocol are aging and, over time, will be increasingly less representative of the appliances recycled through programs. Though several recycling programs have recently been relaunched because of the customer service benefits, energy savings and cost-effectiveness, will ultimately determine the continued deployment of recycling programs. As a result, it is important that future evaluations include metering.

Few metering studies have been conducted since this protocol was first published in April 2013, so there were no new metering data available to update this protocol's regression models. Future metering studies will provide valuable insight into how rapidly per-unit energy savings are declining and, consequently, how long recycling programs can remain viable.

Although it may not be cost-effective to conduct a metering study solely for an appliance recycling program, evaluators can leverage other in-home metering efforts—such as those primarily designed to inform end-use disaggregation; residential heating, ventilating, and air conditioning; or lighting loggers—to collect refrigerator or freezer meter data. This approach may not yield a sufficient sample size for a primary metering study with one effort. However, if the data are collected in accordance with the guidance in Section 4.3.1, collective efforts among evaluators could yield robust samples that are sufficient to update the models included in this protocol.

8 Resources

10 CFR 430.23(A1). (2008). www.gpo.gov/fdsys/pkg/CFR-2011-title10-vol3/pdf/CFR-2011-title10-vol3-part430-subpartB-appA.pdf

ADM Associates, Inc. (April 2008). Athens Research, Hiner & Partners and Innovologie LLC. *Evaluation Study of the 2004-05 Statewide Residential Appliance Recycling Program*. www.calmac.org/publications/EM&V_Study_for_2004-2005_Statewide_RARP_-_Final_Report.pdf

ADM Associates, Inc. (June 2011). *2010 Second Refrigerator Recycling Program; NV Energy – Southern Nevada; Program Year 2010.* https://nwcouncil.app.box.com/s/uljdixyot1i9vcbl44r0ll0xt0xt8w19

Cadmus Group, Inc. (February 8, 2010). *Residential Retrofit High Impact Measure Evaluation Report*. www.calmac.org/publications/FinalResidentialRetroEvaluationReport_11.pdf.

Kema, Inc. (October 24, 2014). *Appliance Recycling Program Impact Evaluation – Volume 1: Report, Work Order 35.* http://www.calmac.org/publications/2010-2012_ARP_Impact_Evaluation_Final_Report.pdf

Khawaja, M. Sami; Rushton, J.; Keeling, J. (2017). *Chapter 11: Sample Design Cross-Cutting Protocol. The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68567. http://www.nrel.gov/docs/fy17osti/68567.pdf

Navigant Consulting. (December 21, 2010). *Energy Efficiency/Demand Response Plan: Plan Year 2 (6/1/2009-5/31/2010) Evaluation Report: Residential Appliance Recycling*. https://www.epa.gov/sites/production/files/2013-12/documents/cee_eval_comedappliancerecyclingpy2evaluationreport2010.pdf

Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Ken Agnew and Mimi Goldberg
DNV GL
Madison, Wisconsin

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68564
November 2017

Contract No. DE-AC36-08GO28308

Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Ken Agnew and Mimi Goldberg
DNV GL
Madison, Wisconsin

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68564
November 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgements

The authors of this chapter wish to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Jessica Granderson of Lawrence Berkeley National Laboratory
- M. Sami Khawaja of Cadmus
- Frank Stern, Carly Olig, and Pace Goodman of Navigant
- Tim Guiterman of Energy Savvy
- Katherine Randazzo and Seth Wayland of Opinion Dynamics
- Rick Ridge of Ridge and Associates
- Kevin Warren of Warren Energy.

Suggested Citation

Agnew, K.; Goldberg, M. (2017). *Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68564. http://www.nrel.gov/docs/fy17osti/68564.pdf

Acronyms

AMI	advanced metering infrastructure
CO	cooling only
DOE	U.S. Department of Energy
HC	heating-cooling
HER	Home Energy Report
HO	heating only
IPMVP	International Performance Measurement and Verification Protocol
IV	instrumental variables
LATE	local average treatment effect
NAC	normalized annual consumption
NOAA	National Oceanic and Atmospheric Administration
NREL	National Renewable Energy Laboratory
RCT	randomized control trial
RED	Random Encouragement Design
SAE	Statistically Adjusted Engineering
TMY	typical meteorological year
UMP	Uniform Methods Project

Protocol Updates

The original version of this protocol was published in April 2013.

This chapter has been updated to incorporate the following revisions:

- Clarified scope for the chapter. This chapter can be used for non-whole-house programs but is not the only approach for whole-house. Clarified application of discussion to daily data without full exploration of options with daily data.
- Expanded allowable modeling options. Included pooled with comparison group, randomized encouragement design, instrumental variables and inverse Mills ratio. Re-worked discussion of and recommendations related to participant-only pooled approach.
- Clarified language (no fundamental changes) around comparison group and net savings.

Table of Contents

List of Tables

1 Measure Description

Whole-building retrofits involve the installation of multiple measures. Whole-building retrofit programs take many forms. With a focus on overall building performance, these programs usually begin with an energy audit to identify cost-effective energy efficiency measures for the home. Measures are then installed, either at no cost to the homeowner or partially paid for by rebates and/or financing.

The methods described here may also be applied to evaluation of single-measure retrofit programs. Related methods exist for replace-on-failure programs and for new construction, but are not the subject of this chapter.[1]

[1] As discussed in the section "Considering Resource Constraints" of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

2 Application Conditions of Protocol

The estimation of the total savings from a multi-measure project requires a comprehensive method for capturing the combined effect of the installed measures. The general method recommended for this type of program is the analysis of consumption data from utility billing records. This method has traditionally been referred to as a billing analysis, and is referred to in this chapter as consumption data analysis.

Unlike the evaluation methods described in most of the other measure-specific chapters of the Uniform Methods Project (UMP), the whole-building analysis methods described in this chapter are designed to provide savings for a program or program segment and do not necessarily produce savings for each participating building. These program-level methods apply only for populations of relatively homogenous buildings. Program-level consumption data analysis as described in this chapter is most commonly applied to residential buildings.

At the individual building level, these methods are consistent with the general approach of International Performance Measurement and Verification Protocol (IPMVP) Option C, Whole Facility. Option C is designed in part to address evaluation conditions that occur with a whole-building retrofit program. However, the IPMVP is designed for individual building analysis, and Option C includes explicit adjustment for non-routine changes, that is, for changes unrelated to the measures of interest that also affect energy consumption. By contrast, the consumption data analysis methods described in this chapter use analysis across multiple buildings to control for non-program-related changes. For a whole-premise program with very heterogeneous participants, such as a large commercial buildings program, the methods described in this chapter are not well suited.

The consumption data analysis approach has strengths and limitations that render it more appropriate to certain types of whole-building program evaluations than to others. This chapter describes how a consumption data analysis can be an effective evaluation technique for whole-building retrofit programs, and it addresses both how and when consumption data analysis should be used.

The evaluation methods noted in this chapter for monthly consumption data are applicable when all of the following are true:

- The whole-building savings from the combination of measures supported by the program are expected to be of a magnitude that will produce statistically significant[2] results given:
    - The natural variation in the consumption data
    - The natural variation in the savings
    - The size of the evaluation sample.

[2] The required level of statistical accuracy may vary from across evaluations. In addition, statistically significant difference from zero is a substantially lower bar than the frequently proposed 90/10 relative precision. Relative precision depends on the magnitude of the savings estimate, the number of participants and, depending on approach, the granularity of the data. Billing analysis results that have 90/50 relative precision are common and may provide acceptable results for the purposes of a program evaluation.

- The baseline for determining savings is the condition of the participating building before the retrofits were made, rather than the standard energy efficiency of the new equipment.
- There is sufficient consumption data available—in the form of monthly or two-month utility billing records—for the participants.
- Consumption data with which to create a comparison group are available for the same timeframe as for the participants for one or more of the following groups:
    - Previous participants—those who took part in the program before the timeframe of the current evaluation
    - Subsequent participants or those who are on a list for future participation in the program
    - Nonparticipants who do not fit either of the first two definitions who are chosen at random or through methods discussed in Section 3.

The evaluation methods described in this protocol are also useful for single-measure programs when all of the requirements listed above are met. Also, note that UMP *Chapter 5: Residential Furnaces and Boilers Evaluation Protocol* uses a consumption data analysis result and addresses the "standard efficiency" baseline issue described in the second bullet above.

2.1 Protocol Applicability to Interval Consumption Data

The methods discussed here are presented as applicable to monthly consumption data. Advanced metering infrastructure (AMI) interval data are now available from many billing systems. These data are commonly available at the hourly or even 15-minute level. From the perspective of billing analysis evaluation, such data, especially when worked with at the daily level, are a finer-grained form of the same basic data. The monthly methods discussed here use, as a dependent variable, average daily consumption as developed from the monthly data. The models will also work when applied to actual daily data. This step is conceptually simple and considerably increases the number of data points available for a single building in a year.

These finer-grained data that are available move consumption data modeling into a wider realm that is beyond the scope of this protocol as initially defined. In addition to likely improvements from more nuanced models that leverage the additional degrees of freedom and more direct relationship between consumption and weather, there are challenges to using daily data. One concern is the increased serial correlation in the modeling process with the more granular data. Hourly data open up a still wider array of issues as the diurnal patterns combined with the unique thermal dynamics of each building demand more complex statistical treatment. A great deal of exploratory work has been done, primarily in the commercial space where interval data have been available for longer. A future protocol will address these issues as methods are consolidated.

3 Savings Framework

Energy consumption data, with their wide availability and their explicit tracking of consumption changes over time, appear to offer a straightforward approach to measuring energy savings. The potential for the change in consumption between two periods to be linked to a program implementation or treatment is compelling. A primary challenge is isolating the effect of the treatment from other sources of change over the same period.

3.1 Components of Change in Consumption

An observed change in consumption between pre- and post-installation periods includes the effect of the whole-building intervention itself, along with the effects of other factors unrelated to the program that may occur in the same timeframe. These effects could include changes in occupancy, physical changes to structure, behavioral changes, weather, etc. Without special attention, these non-program effects may be conflated with program effects leading to incorrect estimates of program effects or savings. This chapter focuses on techniques that attempt to address these concerns including regression techniques, comparison groups, and regression techniques with comparison groups.

3.1.1 Savings Components Captured

The participants' change in consumption includes both direct and indirect effects of the program, in addition to the non-program effects. The program-related effects include the following:

- The **direct effect** of the program measures on the affected systems. For example, replacing an existing light bulb with a more efficient light bulb, while using the lights for the same amount of time, reduces electricity used for lighting.
- **Physical interactive effects** between the directly affected system and other systems in the premise. For example, reduced energy use for lighting also reduces the need for cooling in the summer and increases the need for heating in the winter.
- **Take-back or rebound effects,** where a system is used more *because* it has been made more efficient. For example, the household might pay less attention to turning lights off because the cost of extra lighting use is lower.
- **Participant same-year spillover effects**, where participants install additional energy-savings equipment *because* of the program but *outside* of it, within the post-installation study period. For example, a positive experience with program-provided efficient lighting might lead to installing efficient lighting in other places, or to taking on additional non-lighting efficiency upgrades learned about or encouraged by the program.

The billing analysis does not separate these effects. All are included in the savings estimate captured by the analysis. To the extent takeback or spillover are delayed responses to the measure installation, the measured savings might include only a partial year of these effects. Participant spillover that occurs beyond the timeframe of the post-installation period studied is not captured at all.

Conversely, to the extent that nonparticipants have undertaken measures as a result of the program but outside it, the billing analysis does not capture that nonparticipant spillover as part of the savings. To the contrary, nonparticipant spillover that occurs within the same timeframe as the participant installations may reduce the estimated savings due to the program, since the effect will be treated as part of the exogenous change as captured by a comparison group.

The timeframe of spillover is an important consideration in understanding what the billing analysis results provide. Nonparticipant spillover from past programs that affect the program-year actions of both current-year participants and current-year nonparticipants can be seen as part of the current-year market condition. The billing analysis is capturing savings relative to that current market condition. Spillover or market effects from prior years need to be addressed by different methods than those described here. Spillover to future years outside the analysis period also has no effect on the billing analysis and is not addressed by it.

Contemporaneous nonparticipant spillover due to the current year program and occurring within the same time frame may count against the program in the billing analysis, to some extent. How much the estimated savings is biased downward by the nonparticipant spillover depends in part on the timing of the spillover within the study period, relative to the timing of participation. The bias will be on the order of twice the average contemporaneous nonparticipant spillover savings per household, or less depending on timing.

3.1.2 Free-Ridership as a Component of Savings

The prior two sections highlight the challenge of isolating the program effects of interest (savings) and understanding what components make up those savings. A final aspect of savings that will be addressed throughout the remainder of this chapter is the presence of free-ridership or the difference between net and gross savings. There is a separate UMP chapter that addresses the challenge of free-ridership across evaluation generally (Violette 2017); however, because the choice of consumption data analysis approach has implications for the assessment of free-ridership, it is necessary also to address the subject here.

Chapter 21 of the UMP defines free-ridership as "the program savings attributable to free-riders (program participants who would have implemented a program measure or practice in the absence of the program)." That chapter also provides the following definitions of gross and net savings:

- **Net savings:** The difference in energy consumption *with the program in place* versus what consumption would have been *without the program in place*.
- **Gross savings:** The difference in energy consumption *with the energy efficiency measures promoted by the program in place* versus what consumption would have been *without those measures in place*.

The consumption data analysis approach has implications for whether savings estimates are gross, net or somewhere between.

3.2 Comparison Group Specification

Comparison groups play an important role facilitating the isolation of program effects across a range of disciplines. A comparison groups offers a proxy counterfactual against which an effect can be estimated. In some disciplines a comparison group is used to support an estimate of treatment effect even when data from before and after treatment is not available. The combination of comparison groups with pre- and post-treatment data have the potential to support more robust results than either approach on its own. While there is discussion in Section 4 of approaches that forego comparison groups completely, most consumption data analysis includes some form of implicit or explicit comparison group.

A comparison group consisting of general nonparticipants drawn from the eligible population will control to some extent for factors in the market affecting all customers, such as changes in the economy, prices of energy and of energy-using equipment, or weather.[3] However, these exogenous changes affect different types of customers differently. These different responses to the external factors stem from both the different physical structure and equipment, and the different customer behavioral and decision characteristics. Thus, it's important for the comparison group to match the participants as well as possible in terms of both physical and behavioral/decision-making characteristics.

There are several possible ways to develop a comparison group that is well matched to the participant group in terms of these two dimensions:

1. Starting from a pool of similar customers, randomly assign some customers to receive the whole-building treatment and the others not to. This randomized control trial (RCT) approach is ideal for unbiased measurement of the treatment effect, but for most programs is unrealistic and inconsistent with the program delivery mechanism and theory.
2. Use future program participants as a comparison group for those who participated in the current program year. If the program and participant mix are stable over time, future participants will be similar to current participants, apart from the participation itself.
3. Use past program participants as a comparison group for those who participated in the current program year. The concept is similar to the use of the future participants as a comparison group, as explained further below.
4. Use a set of nonparticipants chosen to match the participants on observable characteristics. Matching characteristics can include consumption in earlier periods, demographic information known from customer records, geography, or explicit average census variables determined from geography. Self-selection bias is still a concern with matched comparison groups, as described below.

3.2.1 Randomized Controlled Trial Control Group

The optimal evaluation scenario for a consumption data analysis is an RCT experimental design. This is the standard approach used across the experimental sciences to (1) isolate treatment (program) effects and (2) establish a causal link between the treatment and the effect.

While a control group constructed by random assignment via an RCT is the "gold standard" for a comparison group, this approach is not practical for most programs. For an RCT, a pool of eligible participants is randomly assigned to one of two groups before the program engagement. This random assignment process assures that the two groups—treatment and control—are similar on average in every respect except for the offer of program treatment. In this context, eligibility

[3] While weather-related change is a form of exogenous change, it is controlled for as well as possible by the weather-normalization process of the Stage 1 models. To the extent that weather is incompletely normalized, the inclusion of a comparison group that has also been weather-normalized will control for any remaining uncontrolled for weather effects.

requirements are defined by those running the program and setting up the RCT (top quartile of consumption, etc.). RCT results will be applicable for the treatment group and other customers that meet eligibility requirements who receive the same treatment (external validity).

The basic analysis structure commonly applied with an RCT is a "difference of differences" calculation. The program-related change is estimated as the difference between the treatment group pre-post difference and the control group pre-post difference.

- For the treatment group, the pre-post difference includes the program-related change plus exogenous change.
- For the control group, the pre-post difference includes only exogenous change.

Because of the random assignment, the average exogenous change is expected to be the same for the treatment and control groups. The control group estimate of exogenous change is used to adjust the treatment group, removing or controlling for that exogenous change. The adjustment is additive and may be positive or negative depending on the direction of the exogenous trend. The final result is an estimate of the treatment group's program-related change.

In the context of energy efficiency programs, true RCT is rare outside of certain types of behavioral programs.[4] The approach, however, provides a good illustration of the ideal characteristics of a comparison group. In particular, the RCT scenario provides an example where the resulting savings estimates are net savings with the effects of free-ridership controlled for. In non-RCT design, the results are generally not net savings. This issue is discussed in Section 3.2.2.

A Random Encouragement Design (RED) approach offers a more flexible approach to incorporating random assignment and addressing challenges related to net savings. We discuss this approach in Section 3.5 after the full range of challenges has been described.

3.2.2 Non-Randomized Comparison Group Development

Where program delivery is not designed as an RCT or other random assignment design, a comparison group is developed after the fact in a quasi-experimental design framework. For that design framework, the term "comparison group" denotes groups that are not randomly assigned but still function similarly to experimental control groups.

The comparison group, which is designed to be as similar as possible to the treatment group during the pre-evaluation period, can be matched to the treatment group using a variety of known characteristics such as geography and pre-program consumption levels.[5] In this context, the

[4] There are multiple reasons why RCT has not been more widely employed. RCT requires denying or delaying participation to a subset of the eligible, willing population and could involve forcing services on people who either do not want them or may not use them. Regulators generally do not support providing tangible services to some customers and not others, outside of limited pilot situations. RCT works for behavior/information programs because there's no forced interference with the premise (recipients can opt out with the utility or effectively opt out by ignoring the reports), and there's no tangible service restricted to only the treatment group.

[5] Since the original writing of this protocol, matched comparison approaches have gained wide acceptance for certain kinds of programs where savings are expected to be small and an RCT control group is not available. Opt-in behavior programs are an example of this kind of program. The limitations of the approach are recognized but no

eligible population from which a comparison group should be drawn follows the requirements for program participation, as much as possible. For example, if participants are required to be single-family dual-fuel households from the utility territory, then the eligible population for a comparison group would start from that definition if that information is available for the general population. As with the true experimental control group, the comparison group is intended to exhibit all of the exogenous, non-program-related effects due to the economy and other factors affecting energy consumption. Thus, the comparison group provides an estimate of exogenous change to use in adjusting participant pre-post impacts.

Unfortunately, matching a comparison group to the treatment group on known characteristics does not produce a true control group. Most importantly, post-hoc matching does not address the issue of self-selection. By the very decision to self-select into a program, the members of the treatment group are different from those of any comparison group that can be constructed post-hoc from nonparticipants.

In theory, many important characteristics can be controlled for in matching or screening to construct the comparison group; however, in reality, the available characteristic data on the customer population is relatively sparse. Also, some important characteristics―such as environmental attitudes―are effectively unobservable. The result is a potential bias that cannot be quantified.

In the context of an energy efficiency program evaluation, the issue of self-selection is complicated by the added dimension of free-ridership. A key characteristic on which we'd like the comparison group to match the participants is whether the customer would adopt the energy efficiency activity in the absence of the program. This characteristic, being a "natural adopter," is unobservable for both participants and nonparticipants. Even for customers who match closely on observable characteristics, those who would adopt on their own are more likely to join a program than those who would not. As a result, self-selection affects the ability to obtain an unbiased estimate of savings, and it affects whether that estimate of savings is best considered gross, net, or something in between.

3.3 Practical Match Comparison Group Development

In some cases, it is not practical to use past or future participants as a comparison group, nor to conduct a pooled[6] consumption data analysis with participation staggered across a year or more. This tends to be the situation when one or more of these conditions are present:

- The program has not been stable over previous and subsequent years.
- The program has not had consistent data-tracking over a sufficient length of time.
- The program participation effects extend over a long time after the tracked participation date, e.g., multiple installation dates, or delayed effects as from a behavioral intervention.

alternative exists. As a result of this work, matching methods such as propensity score matching and minimum distance algorithms have seen wide usage. The specifics of these approaches are beyond the scope of this protocol.

[6] Through this chapter we use the term "pooled" to refer to time-series, cross-sectional data and models.

- The program roll-out results in all participation occurring during only a few months of the year. In such a case, the pooled method will not be useful unless multiple years of participation can be included in the model.

In these cases, either a two-stage or pooled model using a matched nonparticipant comparison group is recommended. One condition for using the general eligible nonparticipant population as a comparison group is that the characteristics of the nonparticipants should be generally similar to those of the participants. Typically, this is not the case. Thus, when participants are different—on the whole—from nonparticipants, a matched group of eligible nonparticipants provides a better comparison group to control for non-program factors among similar premises. However, a matched nonparticipant group is still subject to the same kinds of biases related to naturally occurring savings, self-selection, and spillover, as described above for the general eligible nonparticipant population.

One type of matching stratifies the participants and the comparison pool by observable characteristics, and then randomly selects comparison cases for each stratification cell, proportional to the number in the participant group. Thus, once the matching variables and their ranges or levels are decided, the process is (1) determining the proportion of the participant population in each cell and (2) selecting a nonparticipant sample with the corresponding proportions, from those customers who satisfy the basic eligibility requirements. The following matching factors may be used, depending on their availability:

- Consumption level or other size measure
- Demographics, especially income and education
- Dwelling unit type
- Geography (ZIP code, if feasible)
- Energy end uses.

Another form of matching assigns one (or a specified multiple) specific matched comparison customer(s) to each participant. Propensity score matching and minimum difference algorithms can be used to develop such matching at the customer level across the population or within strata. A variety of approaches for matching are available and new approaches are being tested (Machine learning (e.g., random forests), etc.).

3.4 Self-Selection and Free-Ridership

Whenever a comparison group is selected from customers who were eligible to join the program but chose not to, the potential for self-selection bias is a concern. That is, customers who chose to participate in the program (at a particular time) may have systematic differences from those who did not, resulting in systematic differences in their (changes in) energy consumption, apart from the effects of the program itself. These systematic differences can lead to bias in the savings estimate. While the comparison group construction can control from some of these differences, there are some key differences it can't control for.

A comparison group of eligible nonparticipants controls, in part, for general factors affecting the market, but the general nonparticipant group may respond differently to these general factors than the participants would have without the program.

Matching on observable characteristics, or explicitly including characteristics variables in a regression model, improves the ability of the comparison group to control for the exogenous factors as they affect the participants. However, such matching can't control for the largely unobservable decision-making factors that led the participants to join when they did and the comparison customers not to.

The interaction between self-selection and free-ridership is best illustrated with an example. A true control group is similar to the treatment group with respect to natural levels of energy efficiency activity. For example, if 5% of a population would have installed an energy-efficient furnace without rebate assistance, then the same percentage of both the treatment and control group populations will exhibit this behavior. In the treatment group, some or all of this 5% will participate in the program. By definition, this set of participants consists of free-riders.

In the RCT scenario, the control group does not have access to the program. The naturally occurring savings generated by the 5% natural adopters of the measure in the control group is part of the pre-post non-program exogenous change. The savings from this 5% of the control group that are natural adopters of the measure will equal (on average) the savings for the 5% natural adopters in the treatment group. This natural-adoption portion of treatment-group savings will thus be cancelled out by the corresponding naturally occurring adoption in the control group in the difference of differences calculation. That is, in a true RCT design, naturally occurring energy efficiency savings—and, in the process, free-ridership—are fully removed from the estimate of program-related savings. The result is a "net" estimate of savings; that is, program savings net of free-ridership.

By contrast, an evaluation using a post-hoc comparison group will not generally produce a net savings result. In a non-RCT program scenario, the 5% of households naturally inclined toward the measure adoption all have the option to opt into the programs. Unlike the even allocation across treatment and control groups in the RCT scenario, the allocation of the non-RCT scenario depends on the rate of strategic behavior by the adoption-inclined population. Customers and contractors inclined toward adopting the measure have little reason not to take advantage of the program. This inclination is likely to lead to higher proportion of natural adopters in the participant population, as compared to the general incidence in the population. This differential proportion of natural adopters then affects in multiple ways the level of savings and free-ridership that will be measured by the consumption data analysis.

- First, the participant group includes a higher proportion of natural energy efficiency adopters than would a randomly assigned treatment group (or the general eligible population), due to self-selection into the program. These natural adopter households that strategically opt into the program increase the free-ridership rate among program participants beyond the natural proportion of natural adopters in the eligible population.
- Second, it follows from this that any comparison group developed after the fact from those who chose not to participate will tend to have a lower percentage of natural energy efficiency adopters than would a randomly assigned control group. To return to the

scenario where 5% of the overall population are natural energy efficient furnace adopters, the reduced presence of natural adopters in the comparison group population (<5%) will not negate the self-selected and oversized presence of natural adopters in the participant group (>5% up to 100%).

- Finally, the concerns regarding self-selection beyond the issue of natural adoption are still present. Related to their natural inclination to adopt energy efficiency, the program participants may exhibit different energy-consumption patterns, and different consumption change patterns than the general population. Matching algorithms can help to match the observable characteristics of the comparison group to the participant group. However, the matching inherently cannot match on non-program-induced consumption changes, which are unobservable. To the extent that participation is related to such changes, matching approaches will not fully address self-selection and any associated biases.

These are the key factors that make it impossible for the matched comparison to fully reflect the non-program changes among the participants. As a result, when comparison group change is netted out of the participant change, the netting will control for some but not all of the naturally occurring measure implementation, leaving an unknown amount of free-ridership in the final savings estimate. The resulting estimate is thus somewhere in between net and gross savings.

In the extreme, all households that naturally install an energy-efficient furnace will purchase through the program, leaving no natural energy efficiency purchasing in the non-program population from which the comparison group is constructed. Under this extreme scenario, the comparison group would only provide an estimate of exogenous change apart from natural measure adoption, and would not control for any natural energy efficiency activity. This savings estimate would retain all of the free-rider savings and, thus, would best be classified as a gross savings estimate.

The general recommendations in this whole-building retrofit protocol address these issues by constructing comparison groups that are composed of customers who have opted into the same program in a recent year—or will participate in the near future (pipeline). This approach avoids concerns related to self-selection bias in two ways. Because they have participated or will participate in the same program, they are similar to the participants being evaluated with respect to energy consumption characteristics.[7]

Just as importantly, because they have just participated (or soon will participate) in the program, these previous and future participants are unlikely to install the program measures on their own during their non-participating years.[8] As a result, a comparison group created from previous and

[7] See Randazzo et al. 2017 for an alternative perspective.

[8] If some program-eligible measures are installed without support of the program during the period of time used for the analysis, then the effects of those outside-program installations would be included with the other exogenous change captured by the comparison group. The participants under evaluation would be expected to install outside the program at a similar rate. Depending on the timing of the outside-program installations relative to the timing of participation, some bias can be introduced in either direction. However, if the outside-program installations are timed similarly for current and future participants, and are spread over something like two or more years prior to participation, the future participants will correctly control for current participant outside installations and bias will be minimal.

future participants may be as similar to current-year participants (apart from the program effect itself) as is possible outside of a random assignment design. Thus, the use of such a comparison group is likely to produce a gross estimate of savings that is less biased with respect to self-selection.

3.5 Random Encouragement Design

A Random Encouragement Design (RED) uses random assignment but is more practical to implement for many programs than an RCT. Under the RED, the eligible pool of customers is randomly assigned either to receive supplemental encouragement to participate in the program or not to receive that encouragement. Supplemental encouragement may consist of higher incentive levels, or expanded outreach.

With the RED, a basic difference-of-differences analysis subtracts the average change in consumption among customers who received the supplemental encouragement from the average change among customers who did not receive supplemental encouragement. The averages are calculated across all the customers in each group, not just program participants. The difference of differences is the average change in consumption associated with incremental encouragement. Dividing this difference by the difference in participation rate between the two groups gives the average change in consumption per incremental participant due to encouragement. This incremental change per incremental participant is known as the local average treatment effect (LATE).

A variant of the difference of difference analysis uses a regression approach with instrumental variables (IV), as described in Section 4.3.3. The simplest form of this regression is equivalent to the difference of differences LATE calculation, and provides the same result. A more informed version uses additional explanatory variables.

Regardless of whether difference of differences or basic IV regression is used, the RED produces net savings for the program of interest only under restricted circumstances. The RED does produce incremental net savings per incrementally encouraged participant. However, this incremental savings per incremental participant is not the same as the savings per participant in the base (no-encouragement) program, and in fact may be very different from the base program's net savings. In particular, we anticipate that customers who participate only with supplemental encouragement are less likely to be free-riders than those who participate in the base program. Thus, the RED with basic IV analysis is likely to overstate net savings per participant for the base program. This approach is likely to give an unbiased estimate of net savings for the base program only if:

1. Free-ridership is minimal—that is, net and gross savings are the same
2. There is no relationship between how much energy a customer will save by participating and their inclination to participate (Goldberg et al. 2017).

Nevertheless, obtaining an estimate that can be regarded as a likely upper bound on net savings may itself be useful.

4 Savings Estimation

4.1 Recommendations by Program Characteristics

The consumption data analysis specification and interpretation depend on both the program structure and the corresponding comparison group specification. For a variety of program characteristics, Table 1 shows how the comparison group can be specified and how the resulting savings should be interpreted. Note that some program structures are best for determining net savings, while others are best for determining gross savings. The "consumption data analysis form" column refers to two-stage and pooled modeling approaches which are discussed at length in sections 4.3 and 4.4, respectively.

Table 1. Program Characteristics, Comparison Group Specifications, and Consumption Data Analysis Structure and Interpretation

Randomized Controlled Trial?	Stable Population?	Comparison group	2-Stage and Pooled with Comparison Group	Gross or Net Savings	Unknown Biases
Yes	N/A	Randomly selected control group	Yes	Net	Spillover from T to C, if it exists
No	Yes	Prior and/or future participants	Yes	Gross	Time-varying Characteristics
No	No	Matched comparison group	Yes	Likely between gross and net	Time-varying Characteristics, Self-selection unaccounted for by matching and same-period NP spillover
No	No	General eligible nonparticipants	Yes, With additional characteristics in the 2nd stage or pooled regression	Likely between gross and net	Time-varying Characteristics, Self-selection unaccounted for by regression and same-period NP spillover

Table 1 provides a rough order of preference for analysis form as program conditions become less ideal. Importantly, each approach has strengths and weaknesses that, in specific evaluation scenarios, might justify choosing an approach from lower in the table.

1. **Randomized controlled trial experimental design.** The RCT scenario is unique in that consumption data analysis form will not affect the unbiasedness of the treatment effects. Pooled models will generally provide additional power and specifically, lagged dependent variable models have become a standard approach in the Home Energy Report (HER) literature. These models are discussed in UMP Chapter 21 (Stewart et al. 2017). HER RCT models are almost always designed to measure actual-weather savings, so this modeling approach also avoids distinctions between two-stage and pooled with respect to weather-normalization.

2. **Not randomized, stable program and target population over multiple years.** Table 1 recommends either a two-stage or pooled model with the comparison group created from prior or future participants. Stability, in this case, refers primarily to changes in eligibility rules or major shifts in the supported measures. Changes in targeting and/or marketing may or may not have similar effects. As discussed, the use of the prior/future participants has the potential to address many of the concerns related to self-selection while delivering an estimate of gross savings.[9] The results from the two-stage and pooled approaches should be similar. The two-stage approach offers the increased flexibility with respect to weather modeling relative to the single, mean weather effect estimated in the pooled model. The pooled approach will provide relatively greater precision.
3. **Not randomized, not stable program.** Table 1 recommends a matched comparison group in either a two-stage or pooled approach. As discussed, the matched comparison group should address self-selection bias with respect to the observable characteristics used for matching but not of the remaining self-selection concerns. The savings estimates from this approach will fall somewhere between net and gross. In general, this makes the match comparison group less desirable than the prior/future comparison group. However, in addition to questions regarding program stability for the prior/future approach, prior/future participants will always be relatively less numerous than the eligible matching group. It may be justifiable to use a matched comparison group in place of or in addition to the prior/future participant comparison group. Generally, these results are treated as gross estimates of savings and a separate free-ridership analysis is required (for example, self-reported) to adjust these savings estimates to net savings estimates.
4. **Not randomized, not stable program without matching.** Table 1 offers the final option of a general population comparison group. This approach is similar to the match comparison group approach but with regression variables accounting for differences between the treatment and the more general comparison group. In theory, this approach could be as effective as the matched comparison group. In practice, the data to control for these differences are not readily available. Furthermore, were such variables available, they could also be used either in the matching algorithm or included in the regression with the matched comparison.

4.2 Full-Year and Rolling Analysis Using Prior or Future Participants as the Comparison Group

There are two primary ways to structure the analysis with past and future comparison groups: full year and rolling.

4.2.1 The Full-Year Specification

The full-year approach, illustrated in Table 2, compares the energy consumption from the full year ***before*** the current program year to the full year ***after*** the current program year. Thus, the comparison group consists of customers who either (1) participated in the year that ended a year

[9] Low income programs are a good example of a program that can be stable over time. In the case of low income programs, there is limited expectation of natural occurring savings activity so gross savings may be assumed to equal net saving.

before the start of the current program year[10] or (2) participated in the year that began a year after the end of the current program year.

For example, if the program year occurs in calendar year 2011, then savings would be calculated as the change from calendar year 2010 to calendar year 2012, and the comparison group would be participants from calendar year 2009 and/or calendar year 2013.

If the future participants are used, the full-year approach cannot be applied until the group for later years is identified. Few programs have substantial pipelines, so if future participants are to be used, it may be necessary to wait until late enough in 2013 to identify sufficient future participants with 2010 and 2012 data for the evaluation.

Table 2. Illustration of Analysis Periods for Full-Year Comparison Group, Program Year 2011

Group	Participation Timing	Analysis Period 1 (Pre)	Analysis Period 2 (Post)	Expected Change Period 1 to 2
Past Participants	2009	Jan 2010 – Dec 2010	Jan 2012 – Dec 2012	Non-Program Trend
Current-Year Participants	2011	Jan 2010 – Dec 2010	Jan 2012 – Dec 2012	Program Savings + Non-Program Trend
Future Participants	2013	Jan 2010 – Dec 2012	Jan 2012 – Dec 2012	Non-Program Trend

4.2.2 The Rolling Specification

Although using the full-year comparison group specification is simple, it requires data from farther back in time. The rolling specification, however, allows data from a more-compressed timeframe to be used, as it uses a rolling pre- and/or post-period across the current program year.

Effectively, for each month of the current program year, this method compares the year ending just before that month with the year that begins after that month. The comparison groups for each month's participation are, therefore, the customers who participated one year before and/or the customers who participated one year later. This structure is illustrated in Table 3 for program year 2011.

[10] Some find it counterintuitive to use past participants for the comparison group because they are no longer similar to pre-program participants by the very fact of their participation. They are, however, assuming a stable program and participation mix, similar in all other ways to post-program participants. The difference-in-differences structure relies on an additive period-to-period change factor that works equally well with past or future participants. Future participants represent how current participants would have changed had they not participated in this year. That is, they capture the effect of all changes other than participation itself. Similarly, past participants represent how current participants would have changed had they already participated prior to this year. Thus, the prior participants also capture the effect of all changes other than participation itself.

Table 3. Illustration of Analysis Periods for Rolling Comparison Group, Program Year 2011

Group	Participation Timing	Analysis Period 1 (Pre)	Analysis Period 2 (Post)	Expected Change Period 1 to 2
Past Participants	Feb 2010	Mar 2010 – Jan 2011	Mar 2011 – Feb 2012	Non-Program Trend
	Jun 2010	Jul 2010 – May 2011	Jul 2011 – Jun 2012	Non-Program Trend
	Dec 2010	Jan 2011 – Nov 2011	Jan 2012 – Dec 2012	Non-Program Trend
Current-Year Participants	Feb 2011	Mar 2010 – Jan 2011	Mar 2011 – Feb 2012	Program Savings + Non-Program Trend
	Jun 2011	Jul 2010 – May 2011	Jul 2011 – Jun 2012	Program Savings + Non-Program Trend
	Dec 2011	Jan 2011 – Nov 2011	Jan 2012 – Dec 2012	Program Savings + Non-Program Trend
Future Participants	Feb 2012	Mar 2010 – Jan 2011	Mar 2011 – Feb 2012	Non-Program Trend
	Jun 2012	Jul 2010 – May 2011	Jul 2011 – Jun 2012	Non-Program Trend
	Dec 2012	Jan 2011 – Nov 2011	Jan 2012 – Dec 2012	Non-Program Trend

The comparison group, which captures exogenous change through the evaluation time span, ultimately provides an average of the exogenous change through the 12 months of the current evaluation year. Thus, this group should be selected in such a way that the estimate of exogenous change across the 12 months will be from pre- and post-data periods that are similarly distributed across the evaluation year as the current participants.

If participation rates are stable across the multiple program years being used, the rolling specification will often accomplish a similar distribution over the year without additional effort. However, when using the rolling specification, examine the pattern of participation within each season over the applicable years for each of the two or three groups (current year and past and/or future participants). If the distribution is not similar,[11] then the comparison group should be properly scaled using ***one*** of these methods:

- On a season-by-season basis, sample from the past and/or future comparison groups in proportion to the current year's participation.

[11] This may indicate changes in the program or the program participants that may affect whether this is, in fact, a valid comparison group.

- Re-weight the past and future participants to align with the current-year participants' timing distribution. That is, for a comparison group customer who participated in season s, assign the weight f_{Ts}/f_{gs} where f_{gs} is the proportion of past or future participant group g who participated in seasons and f_{Ts} is the proportion of the current participant group. Then apply these weights in the second-stage analysis.

Generally, for any set of participant sites, the comparison sites need two years of either all pre- or all post-consumption data that cover the year before and after that installation month. This gives the analyst the freedom to create these comparison group pre- and post- data periods using exactly the same distribution as the current year participant dates.

4.3 The Two-Stage Approach

4.3.1 Stage 1. Individual Premise Analysis

For each premise in the analysis, whether in the participant or comparison group, do these activities:

1. Fit a premise-specific degree-day regression model (as described in Step 1, below) separately for the pre- and post-periods.
2. For each period (pre- and post-) use the coefficients of the fitted model with normal-year degree days to calculate weather-normalized annual consumption (NAC) for that period.
3. Calculate the difference between the pre- and post-period NAC for the premise.

The site-level modeling approach was originally developed for the Princeton Scorekeeping Method (PRISM™) software (Fels et al. 1995). (The theory regarding the underlying structure is discussed in materials for and articles about the software [Fels 1986].) Stage 1 of the analysis can be conducted using PRISM or other statistical software.[12]

4.3.1.1 Step 1. Fit the Basic Stage 1 Model

The degree-day regression for each premise and year (pre- or post-) is modeled as:

Equation 1

$$E_m = \mu + \beta_H H_m + \beta_C C_m + \varepsilon_m$$

where:

[12] PG&E has supported an effort in California called CalTRACK that is designed to document a set of methods for calculating site-based, weather-normalized, metered energy savings from an existing conditions baseline and applied to single family residential retrofits using data from utility meters, to support various use cases including a residential pay-for-performance pilot. The effort references this UMP chapter, primarily related to Stage 1, site-level modeling. The results of that effort were not finalized at the time of this revision but will offer another source of instruction related to the practical technical methods discussed here. http://www.caltrack.org/

E_m = Average consumption per day during interval *m*

H_m = Specifically, $H_m(\tau_H)$, average daily heating degree days at the base temperature(τ_H) during meter read interval m, based on daily average temperatures on those dates

C_m = Specifically, $C_m(\tau_C)$, average daily cooling degree days at the base temperature(τ_C) during meter read interval m, based on daily average temperatures on those dates

μ = Average daily baseload consumption estimated by the regression

β_H, β_C = Heating and cooling coefficients estimated by the regression

ε_m = Regression residual.

4.3.1.2 Stage 1 Model Selection

Fixed Versus Variable Degree-Day Base

In the simplest form of this model, the degree-day base temperatures τ_H and τ_C are each pre-specified for the regression. For each site and time period, only one model is estimated using these fixed, pre-specified degree-day bases.

For ease of processing and of meeting data requirements, the industry standard for many years was to use a fixed 65°F for both heating and cooling degree-day bases. However, actual and normal hourly weather data are easily available now, providing flexibility in the choice of degree-day bases. In general, a degree-day base of 60°F for heating and of 70°F for cooling usually provide better fits than a base of 65°F

The fixed-base approach can provide reliable results if there are only moderate differences between the actual weather used to estimate the models and normal/typical meteorological year (TMY) weather used to construct NAC. When this is the case and data used in the Stage 1 model span all seasons, NAC is relatively stable across a range of degree-day bases. However, the decomposition of consumption into heating, cooling, or baseload coefficients is highly sensitive to the degree-day base. For houses in which the degree-day bases are different from the fixed degree-day bases used, the individual coefficients will be more variable and, potentially, biased as will the combined NAC. As a result, if the separate coefficient estimates will be used for savings calculations or for associated supporting analysis, the fixed degree-day base simplification is not recommended. Similarly, under extreme weather conditions, the variable degree day base is recommended to control for a greater portion of weather-related exogenous change along with a comparison group to address remaining weather-related change.

The alternative approach is variable degree-day, which entails the following steps:

1. Estimating each site-level regression and time period for a range of heating and cooling degree-day base combinations (including dropping heating and/or cooling components).
2. Choosing an optimal model (with the best fit, as measured by the coefficient of determination R^2 or CV(RMSE) within a specification and adjusted R^2, AIC, or BIC across models with different variables[13]) from among all of these models.

[13] Akaike information criteria and Bayesian information criteria are alternative measures for comparing the goodness of fit of different models.

The variable degree-day approach fits a model that reflects the specific energy consumption dynamics of each site. In the variable degree-day approach, the degree-day regression model for each site and time period is estimated separately for all unique combinations of heating and cooling degree-day bases, τ_H and τ_C across an appropriate range. This approach includes a specification in which one or both of the weather parameters are removed.

Degree Days and Fuels

For the modeling of natural gas consumption, it is unnecessary to include a cooling degree-day term. The gas consumption models tested should include the heating only (HO) and mean value options. Gas-heated households having electric water heat may produce models with negative baseload parameters. The models for these households should be re-run with the intercept (baseload) suppressed.

For the modeling of electricity, a model with heating and cooling terms should be tested, even if the premise is believed not to have electric heat or not to have air conditioning. Thus, for the electricity consumption model, the range of degree-day bases must be estimated for each of these options: a heating-cooling (HC) model, HO, cooling only (CO), and no degree-day terms (mean value).

Degree Days and Set Points

If degree days are allowed to vary:

- The estimated heating degree-day base τ_H will approximate the highest average daily outdoor temperature at which the heating system is needed for the day
- The estimated cooling degree-day base τ_C will approximate the lowest average daily outdoor temperature at which the house cooling system is needed for the day.

These base temperatures reflect both average thermostat set point and building dynamics, such as insulation and internal and solar heat gains.

The average thermostat set points may include variable behavior related to turning on the air conditioning or secondary heat sources. If heating or cooling are not present or are of a magnitude that is indistinguishable amidst the natural variation, then the model without a heating or cooling component may emerge the most appropriate model.

The site-level models should be estimated at a range of degree days that reflects the spectrum of feasible degree-day bases in the population. In general:

- A range of heating degree-day bases (from 55°F through 70°F) cover the feasible spectrum for single-family dwellings
- Cooling degree-day bases ranging from 65°F through 75°F should be sufficient.[14] (Note that the cooling degree-day base must always be higher than the heating degree-day base.)

[14] In both cases, it is important to remember that temperatures are based on average daily temperature and will be aggregated over a month or more of time.

A wider range of degree-day bases increases processing time, but this approach may provide better fits in some cases.

Plotting daily average consumption with respect to temperature provides insight into the inflection points at which heating and cooling consumption begin. However, mixed-heat sources may make a simple characterization of heat load such as this difficult.

For each premise, time period, and model specification (HC, HO, or CO), select as the final degree-day bases the values of τ_H, and τ_C that give the highest R^2, along with the coefficients μ, β_H, β_C estimated at those bases. Models with negative parameter estimates should be removed from consideration, although they rarely survive the optimal model selection process.

4.3.1.3 Optimal Models

When the optimal model degree-day bases determined by the R^2 selection criterion are within the extremes of the temperature range tested, identify an optimal model. However, if the best-fitting model is at either extreme of the degree-day bases tested, this may not be the case. An extreme high- or low-degree-day base could indicate that the range of degree-day bases tested was too narrow, or it may reflect a spurious fit on sparse or anomalous data. If widening the degree-day base range or fixing anomalous data does not produce an optimal model within the test range, these sites should be flagged and plotted and the analyst should then decide whether the data should be kept in the analysis.

The practical response to degree-day base border solutions is to default to the fixed degree-day approach. In this case, the fixed degree-day bases could be fixed at the mean degree-day bases of all sites that were successfully estimated with a meaningful (non-extreme) degree-day base. Otherwise use 60°F for heating and 70°F for cooling. The NAC for these fixed degree-day base sites will still be valid, but the heating and cooling estimated parameters for these sites are potentially biased. This approach maximizes the information learned where the variable degree-day base approach works, but it defaults to the more basic approach where it fails.

Apply a consistent reliability criterion based on R^2 and the coefficient of variation (primarily for baseload-only models) to all site-level models. Ranking by R^2 is the simple way to identify the optimal degree-day choice within each specification (HC, HO, and/or CO). Use an appropriate statistical test to determine the optimal model among all of the different specifications (HC, HO, CO, and mean). The simplest acceptable selection rule is as follows[15]:

- If the heating and cooling coefficients in the HC model have p-values[16] less than 10%, retain both.
- Otherwise:
    - If either the heating coefficient in the HO model or the cooling coefficient in the CO model has a p-value of less than 10%, retain the term (heating or cooling) with the lower p-value.

[15] Adjusted R2, AIC or BIC are also used.
[16] A measure of statistical significance.

- If neither the heating nor the cooling coefficient has a p-value of less than 10% in the respective model, drop both terms and use mean consumption.
- For sites with no weather-correlated load or with a highly variable load, the mean usage-per-day may be the most appropriate basis for estimating normal annual consumption.

It is always possible to estimate a "best" model, but a number of caveats—such as those listed here—remain. Any interpretation of the separate heating and cooling terms from either the first stage of the stage-two model or the pooled model must recognize that these other uses are combined to some extent with heating and cooling.

- These models are very simple.
- Many energy uses have seasonal elements that can be confounded with the degree-day terms.
- During cold weather, the consumption of hot water, the use of clothes washers and dryers, and the use of lighting all tend to be greater.
- In summer, the refrigerator load and pool pumps tend to be greater.
- Internal loads from appliances, lighting, home office, and home entertainment reduce heating loads and increase cooling loads.
- Low-e windows and window films increase heating loads and reduce cooling loads.

To review, fixed degree-day base models can be used if the only information derived from the model is normalized annual consumption, because NAC is generally stable regardless of the degree-day base used. ***Fixed degree-day base models should not be used if the separate heating, cooling, or base components are to be interpreted and applied as such.***

4.3.1.4 Step 2. Applying the Stage 1 Model

To calculate NAC for the pre- and post-installation periods for each premise and timeframe, combine the estimated coefficients μ, β_H, and β_C with the annual normal- TMY[17] degree days H_0 and C_0 calculated at the site-specific degree-day base(s), τ_H and τ_C. Thus, for each pre- and post-period at each individual site, use the coefficients for that site and period to calculate NAC. This example puts all premises and periods on an annual and normalized basis.

$$NAC = \mu * 365.25 + \beta_H H_0 + \beta_C C_0$$

The same approach can be used to put all premises on a monthly basis and/or on an actual weather basis. In instances where calendarization may be required, it may be preferable to use this approach to produce consumption on a monthly and actual weather basis, rather than using the simple pro-ration of billing intervals.

[17] Discussed in Section 4.4.6 in UMP *Chapter 17: Residential Behavior Evaluation Protocol*.

4.3.1.5 Step 3. Calculating the Change in NAC

For each site, the difference between pre- and post-program NAC values (ΔNAC) represents the change in consumption under normal weather conditions.

4.3.2 Stage 2. Cross-Sectional Analysis

The first-stage analysis estimates the weather-normalized change in usage for each premise. The second stage combines these to estimate the aggregate program effect, using a cross-sectional analysis of the change in consumption relative to premise characteristics.

Three forms of the stage-two regression are recommended. Influence diagnostics should be produced for all stage-two regressions with outliers removed. Alternatively, some evaluators remove outliers based on data-dependent criteria such as 2.5 inter-quartile ranges from the median percent savings (established separately for the participant and comparison groups because they have different central tendencies and variances).

4.3.2.1 Form A. Mean Difference of Differences Regression

As the most basic form of the stage-two regression, this approach produces the same point estimates as taking the difference of the average pre- and post-differences; however, it will produce slightly different standard errors as it assumes a common variance.

Equation 2

$$\Delta NAC_j = \beta + \gamma I_j + \varepsilon_j$$

where:

ΔNAC_j = change in NAC for customer j
I_j = 0/1 dummy variable, equal to 1 if customer j is a (current-year) participant, 0 if customer j is in the comparison group
β, γ = coefficients determined by the regression
ε_j = regression residual.

From the fitted equation:

- The estimated coefficient γ is the estimate of mean savings.
- The estimated coefficient β is the estimate of mean change or trend unrelated to the program.

The coefficient β corresponds to the average change among the comparison group, while the coefficient γ is the difference between the comparison group change and the participant group change. That is, this regression is essentially a difference-of-differences formulation and can be accomplished outside of a regression framework as a difference of the two mean differences.

4.3.2.2 Form B. Multiple Regression with Program Dummy Variables

This form allows for the estimation of savings for different measures or groups of measures. It may also include other available premise characteristics that can improve the extrapolation of billing analysis results to the full program population.

For whole-building programs, the typical savings magnitude can vary substantially across the different measures that may be implemented under the program. Regression with a single dummy variable produces a single average savings per premise across premises. With widely varying actions across premises, this average may not be well determined. Allowing for different average savings for different measure groups can result in a better estimated model. However, it's typically not possible to isolate the effects of each individual measure. It's most effective then to include only a handful of measure groups, such as one to three large-impact measures individually, plus all others as a group.

Equation 3

$$\Delta NAC_j = \Sigma_q \beta_q x_{qj} + \Sigma_k \gamma_k I_{kj} + \varepsilon_j$$

where:

I_{kj}	= 0/1 dummy variable, equal to 1 if customer j received measure group k in the current year, 0 if customer j is in the comparison group and/or did not receive measure group k.
x_{qj}	= value of the characteristics (square footage, number of occupants, etc.) variable q for customer j. Let x_{0j}, the first term of this vector, equal 1 for all premises, so that β_0 serves as an intercept term.
β_q, γ_k	= coefficients determined by the regression.

From the estimated equation:

- The estimated coefficient γ_k is the estimate of mean savings per participant who received measure group k.
- The coefficient β_q is the estimate of mean change or trend unrelated to the program per-unit value of variable x_q.

This form may be used with any of the following:

- Multiple characteristics variables x_q and a single measure dummy variable I
- Multiple dummy variables I_k and a single characteristics variable x (other than the intercept)
- Only an intercept term (no premise characteristics) and a single dummy variable, I.

If only an intercept term and a single dummy variable are used, this form reduces to the first model type. For this type of regression to be meaningful, it is essential that the characteristics variables (x_q) are obtained in a consistent manner for both the participants and the comparison group. For many programs, if the comparison group is future or prior participants, these variables may be obtained from tracking data collected the same way across the program years.

4.3.2.3 Form C. Statistically Adjusted Engineering Regression with Program Dummy Variables

This form adds the expected savings into the regression specification. If the expected savings from the tracking data are more informative than the simple indicator variable used in the previous specifications, then this approach should have greater precision. The model structure assumes an additive relationship between multiple measures which may not reflect interactive effects. Measure combinations can be parameterized to capture interaction effects explicitly.

Equation 4

$$\Delta NAC_j = \Sigma_q \beta_q x_{qj} + \Sigma_k \gamma_k I_{kj} + \Sigma_k \rho_k T_{kj} + \varepsilon_j$$

where:

T_{kj}	= tracking estimate of savings for measure group k for current-year participating customer j, 0 for customer j in the comparison group
β_q, γ_k , ρ_k	= coefficients determined by the regression

From the fitted equation:

- The mean program savings must be calculated using the coefficients on both the participation dummy variables and the tracking estimates of savings. That is, the estimated mean program savings for measure group k with mean tracking estimate T_k is:

  $$S_k = \gamma_k + \rho_k T_{k\_}$$

- The coefficient β_q is the estimate of mean change or trend unrelated to the program per-unit value of variable x_q.

This form may be used with any of the following:

- Multiple characteristics variables x_q and a single measure group
- Multiple measure groups k and a single characteristics variable x (other than the intercept)
- Only an intercept term, no premise characteristics and a single measure group.

For each measure group k in the model, both the dummy variable I_k and the tracking estimate T_k should be included.

A simpler Statistically Adjusted Engineering (SAE) form that omits the participation dummy variable has the nominal appeal of the coefficient ρ_k being interpreted as the "realization rate," the ratio of realized to tracking savings. However, inclusion of the tracking estimate without the corresponding dummy variable can lead to understated estimates of savings due to errors from omitted variables bias.

If the tracking estimate of savings is a constant value for all premises, the inclusion of the tracking estimate will not improve the fit. Moreover, if the tracking estimates vary but in ways that are not well correlated with actual savings, the fit will tend to be poor, with some savings coefficients not significant and others not realistic. As for the multiple-dummy variable approach Form B described in Section 4.3.2.2, the SAE approach works best if the number of separate measure groups k is kept small. If the SAE approach does not produce meaningful results, the multiple- or single-dummy-variable version is preferred.

4.3.3 Instrumental Variables Regression

Instrumental variables (IV) regression addresses a potential bias in the basic regression that can arise if the tendency to participate is correlated with the change in consumption unrelated to the program. Such a correlation will tend to exist if any of the following are true:

- Free-ridership is present, at a non-negligible level
- The comparison group includes customers for whom the program measures aren't applicable
- Customers tend to participate in the program at times when they're taking other actions or have life events that generally tend to increase (or that generally tend to decrease) consumption.

Measure applicability is a particularly a concern when free-ridership is present. Customers for whom the program measures wouldn't apply or wouldn't make sense have zero natural adoption and don't participate in the program. Thus, the inclusion in the comparison group of customers who couldn't benefit from the program measures exacerbates the mismatch between the participant and comparison groups' rates of natural adoption.

The IV method adds an additional step to the regression process. Specifically, a model that predicts participation as a function of observable variables is fit. If an RED is used, the encouragement dummy variable becomes a predictor in the participation model. Common forms of the participation model include a logit or probit.

The fitted model is then used to calculate the participation probability for each customer in the analysis, and this participation probability is substituted for the participation dummy in Eq. 2 or 3. In the simplest form with an RED, the encouragement variable is the only predictor in the participation equation, and Eq. 2 with the substitution of predicted for observed participation is used for the analysis. In this form, the result is equivalent to the difference of differences LATE estimator described in Section 3.5.

Conditions for the participation model specification include the following:

1. It should include all the explanatory variables x_q included in Eq. 3 above.
2. It should include one or more variables that DON'T directly affect energy consumption but DO affect participation.
3. If there are any additional (observable or unobservable) consumption drivers that are left out of the consumption equation, the participation model predictors must be unrelated to any of these omitted terms.

The IV approach may be used with or without an RED. However, without an RED, it is difficult for the 2^{nd} participation model condition to be satisfied. It also may be difficult to get good predictive power for the model. If the participation model has weak ability to separate high- from low-participation customers, the IV analysis will tend to yield savings estimates with high variance.

The basic IV analysis cannot provide an unbiased net savings estimate for the main program when free-ridership is present. However, the IV analysis with RED does control for unobservable factors that affect naturally occurring change but don't also affect the net savings a customer will have if they join the program. In many whole-building programs there is a tendency to join a program at a time of major renovation, which tends to increase consumption. On the other hand, customers might choose to join a pay-for-performance program if they anticipate household changes that will tend to bring consumption down. The RED can eliminate the bias due to factors such as these that tend to work in a particular direction for a particular program. If the three conditions noted above for the participation model are met, the average effect of such factors, for a given participation probability, is zero. As a result, there is no confounding of these unrelated changes with the estimated participation effect.

Moreover, as noted in Section 3.5, when free-ridership is present, the LATE estimate from the RED with basic IV analysis does give the net savings per incremental participant attributable to the incremental encouragement. Since free-ridership is likely to be lower among those who require supplemental encouragement to join, this LATE estimate can be regarded as an upper bound on the base program net savings.

The use of RED and IV methods is discussed more fully in Goldberg et al. (2017). That work also describes an additional method that can potentially provide an unbiased estimate of net savings for the main program, using an extension of the basic IV method. While that method is promising, further empirical work is needed before specific recommendations can be offered for its use.

4.3.4 Choosing the Stage-Two Regression Form

The mean difference-of-differences regression estimate (described earlier) is recommended if the following three conditions are met:

- Only overall average program savings is to be estimated, rather than separate savings for different groups of measures
- Factors that may be associated with differences in the magnitude of the non-program trend (such as square footage) are the same on average for the current-year participant group as for the comparison group
- More precise estimates are not required, or additional data that could yield a more accurate estimate are not available.

The second general model, Form B (Multiple Regression with Program Dummy Variables), is recommended if:

- Either (a) separate savings estimates are desired for different groups of measures, ***or*** (b) factors that may be associated with differences in the magnitude of the non-program trend (such as square footage) are not the same on average for the current-year participant group as for the comparison group

- Informative tracking estimates of savings are not available.

The third general model, Form C (SAE Regression With Program Dummy Variables), which incorporates a tracking estimate of savings, is preferred when there are both an informative tracking estimate of savings ***and*** an interest in more refined estimates than can be obtained with the simplest model version.

Forms B and C make it possible to extrapolate the consumption data analysis results back to the full tracking data based on measure-level results. This may be of particular importance, depending on the extent and nature of the attrition of tracking data sites out of the analysis dataset.

If an informative tracking estimate is not available but there are characteristics variables likely to correlate with savings, then a proxy for savings constructed from these characteristics variables can be substituted for the tracking estimate. Proxies that may usefully inform a second-stage model include count of light bulbs and the square footage of installed insulation.

4.4 Pooled Fixed-Effects Approach

The pooled approach can be specified either with a comparison group or with multiple years of participants. With a comparison group, the pooled model is a pooled version of the 2-stage approach discussed above. With multiple years of participants included in the pooled model the later participants are implicitly performing the function of comparison group. The comparison group approach offers a more straightforward specification and is the focus of this section.[18]

The basic structures of the site-level and the second-stage consumption data model are effectively combined in the pooled approach. All monthly participant and comparison group consumption data (both pre- and post-installation) are included in a single model. This model has:

- A site-level fixed-effect component (analogous to the site-level baseload component)
- A monthly fixed effect
- A participant group indicator variable (absorbed into the site-level fixed effect when not interacted with other variables)
- A post-installation indicator variable capturing the change in the post-installation period across participant and comparison groups
- A participant-post combined indicator that captures the savings estimate
- Heating and cooling components interacted with the participant indicator variable, the post-installation indicator variable, and the participant-post combined indicator variable.

4.4.1 Recommended Form of Pooled Regression

An example pooled model equation is as follows:

[18] The discussion in the section parallels discussion in Section 4.3.6 of *Chapter 17: Residential Behavior Protocol*. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures.

Equation 5

$$E_{jm} = \mu_j + \phi_m + \beta_H T_j H_{jm} + \lambda P_{jm} + \lambda_H P_{jm} H_{jm} + \gamma T_j P_{jm} + \gamma_H T_j P_{jm} H_{jm} + \varepsilon_{im}$$

where all variables have already been defined except for these:

μ_j	=	Unique intercept for each participant j
ϕ_m	=	0/1 Indicator for each time interval *m*, time series component that track systematic change over time
P_{jm}	=	0/1 Indicator variable for the post-installation period for both treatment and comparison groups.
$\beta_H, \lambda, \lambda_H, \gamma, \gamma_H$	=	coefficients determined by the regression

This specification only includes heating terms (H_{jm}), as if for a gas analysis; however, analogous cooling terms should be included for an electric pooled model.

The parameter interactions that include only the variable P_{jm} capture the post-period effect for both participants and the comparison group. The parameter interactions that only include I_j control for differences between the participant group and comparison group in the pre-period.[19] The parameter interactions with both P_{jm} and I_j represent the post period effect on participants given the other interactions. This specification is the regression version of the difference in difference approach.

The mean program savings is calculated using the following equation:

$$S = \gamma * 365.25 + \gamma_H H_0$$

where:

H_0 = TMY degree days at the base for the regression

The pooled regression can also be specified as an SAE model.

4.4.2 Choice of Pooled Form

The pooled approach features a simplified weather-normalization structure compared to the site-level modeling in the two-stage approach. All buildings are characterized by a mean heating and/or cooling slope calculated from a fixed degree day base. In addition, the panel structure requires regression errors to be clustered at the building level to address the lack of independence of consumption across month within a building. The primary advantage of the pooled structure is the avoidance of site-level modeling altogether. In general, the pooled approach will also provide estimates with a higher precision, even after clustering, due to the increased size of the dataset.

4.5 Data

4.5.1 Basic Data Preparation

Before a consumption data analysis can be performed, the following activities must be done. The details of these steps are provided later in this section.

[19] The mean difference between the two groups is accounted for in the site-level fixed effect, μ_j.

1. ***Obtain program tracking data for current year participants.*** The tracking data should identify what program measures were installed and on what date. These data may also include some customer or building characteristics.
2. ***Identify the comparison group customers.*** Obtain tracking data for these customers if they are previous or future participants, so as to assure that all comparison group consumption data are either fully pre- or fully post-participation in the program.
3. ***Obtain consumption data files from billing records for each building in the analysis.*** This may require mapping participant account numbers to premise accounts. Buildings with occupant turnover during the evaluation period should be assessed separately and may warrant removal from the analysis.
4. ***Screen and clean the consumption data*** as described in *Data Requirements and Collection Methods,* Section 4.5.2.
5. ***Convert the billing records for each meter reading interval*** to average consumption-per-day for each premise.
6. ***Identify the pre- and post-periods for each premise in the analysis.*** Based on the installation dates, the pre- and post-installation periods are defined for each participant to span approximately 12 months before and approximately 12 months after installation. The billing interval or intervals during which the measure was installed for a particular participant include both pre- and post-installation consumption days. These transitional billing intervals should be excluded from the analysis. (The excluded billing intervals are referred to as the blackout intervals for that participant.) The post period is identified with 0/1 dummy variable.
7. ***Identify the nearest weather station associated with each premise in the analysis.*** The utility may maintain a weather station look-up for this purpose, so use that if it is available. In general, weather station assignments should consider local geography rather than simply selecting the nearest station. For example, in California, the weather station should be in the same climate zone as the home. Also, consider all significant elevation differences in the station assignment.
8. ***Obtain daily temperature data from each weather station*** for a period that matches the consumption data.
9. ***Determine for each weather station the actual and normal heating and cooling degree days*** for degree day base temperatures—from 55^{o}F through 75^{o}F—for each day included in the analysis, as is detailed in the *Data Requirements and Collection Methods* section below.
10. ***Calculate average daily degree days*** for the exact dates of each bill interval in the consumption data.

4.5.2 Data Requirements and Collection Methods

A consumption data analysis requires data from multiple sources:

- Consumption data, generally from a utility billing system
- Program tracking data
- Weather data.

This section describes the required data for a whole-building retrofit billing analysis and the steps for using these data correctly.

4.5.2.1 Consumption Data

The consumption data used in a consumption data analysis are generally stored as part of the utility billing system. Because these systems are used by evaluators relatively infrequently, recovering consumption data from the system can be challenging. To obtain the needed data, prepare a written request specifying the data items, such as:

- Unique site ID
- Unique customer ID
- Read date
- Consumption amount
- Read type (indicating estimated and other non-actual reads)
- Variables required to merge consumption data with program tracking data
- Location information or other link to weather stations
- Customer tenancy at the premise (the tenancy starting and ending dates)
- Other premise characteristics available in the utility customer information system, including dwelling type, heating or water heating fuel indicators, or participation in income-qualified programs.

It is essential to establish the unique site identifier with the help of the owner of the data at the utility. Note that the unique site ID specifies the unit of analysis. Usually, a combination of customer and site/premise ID identifies a particular location with the consumption data for the occupant.

The primary data used for a consumption data analysis are the consumption meter reads from the utility revenue meter, and these readings are typically taken monthly or bimonthly for gas and electric utilities in the United States. The consumption data are identified with specific time intervals by a meter read date and either a previous read date or a read interval duration. Average daily consumption for the known monthly or bi-monthly time interval is calculated by combining these data, which then serve as the dependent variable for all of the forms of consumption data regression.

The remaining requested variables serve one of three purposes:

- Linking the consumption data with other essential data sources (such as program tracking data and weather data)
- Providing information that facilitates the cleaning of the consumption data
- Providing data for characterizing the household so as to improve the quality of the regression models.

Consumption Data Preparation

Consumption data received from the service provider are likely to be subject to some combination of the following issues, which are provided here as a checklist to be addressed. It is almost impossible to prescribe definitive rules for addressing some of these issues, as they arise from the unique conditions of each billing system. This list represents the common issues encountered in consumption data and provides basic standards that should be met. The general goal should be to limit the analysis to intervals with accurate consumption data with accurate beginning and ending dates.

- ***Zero reads.*** Zero electric reads are rare and usually indicate outages, vacancy, or other system issues. Zero gas reads, however, are more common. Infrequent zeros in an electric data series can be ignored, as can zero reads in gas series during the non-heating months. Sites with extensive electric zero reads or zero gas reads during the heating season should be identified and removed.
- ***Extreme data.*** Sites with extreme reads should be removed unless evidence indicates that high-level usage patterns are typical. Atypical extreme spikes are frequently the result of meter issues, so it is best to omit them from the analysis. For smaller populations: (1) Plot and review consumption levels above the 99^{th} percentile of all consumption levels. Alternatively, flag points that are more than three inter-quartile ranges away from the median consumption. (2) Develop realistic consumption minima and maxima for single-family homes. The decision rule should be applied consistently to the participant and comparison groups.
- ***Missing data.*** Missing data should be clearly understood. Some instances are self-explanatory (pre- or post-occupancy), but many are not, and these require an explanation from the utility data owner. Because true missed reads are generally filled with estimations, missing data in the final consumption indicate an issue worth exploring.
- ***Estimated reads.*** A read type field, available from most billing systems, indicates whether a consumption amount is from an actual read or some form of system estimate. Any read that is not an actual read should be aggregated with subsequent reads until the final read is an actual read. The resulting read will cover multiple read intervals, but the total consumption will be accurate for the aggregated intervals.
- ***First reads.*** The first read available in a consumption data series may correct for many previous estimated reads. Each site data series used for the analysis should begin with a consumption value that is a confirmed single-read interval. This entails removing all leading estimated reads from the series and then removing one additional, non-estimated leading read from each site data series.
- ***Off-cycle reads.*** Monthly meter reading periods that span fewer than 25 days are typically off-cycle readings, which typically occur due to meter reading problems or changes in occupancy. These periods should be excluded from the analysis.
- ***Adjustments.*** Adjustment reads may either be single reads that are out of the normal schedule or reads combined with a normally scheduled read. Adjustments may be indicated by the read-type variable, or they may appear, for instance, as a consistent spike in December reads. Adjustments correct a range of errors in previous consumption data in a one-time, non-informative way. Unless the magnitude of the adjustment is small, such adjustments necessitate the removal of prior data from a site and may require the complete removal of the site if enough data are compromised.

- ***Overlapping read intervals.*** Because overlapping read intervals may indicate an adjustment or a data problem, they should be discussed with the data owner. If these read intervals undermine the consumption-weather relationship, then the site must be removed.
- ***Multiple meters.*** Although having multiple meters is rare in single-family housing, this situation does exist. When multiple meters are read on the same schedule, as is usually true for such residences, the meter reads for the same home should be aggregated to the household level for each meter reading interval.

As consumption data analysis is generally applied to the full population of a program, dropping small percentages of sites is unlikely to affect the results. However, if the number of removed sites increases beyond 5%, it is worth considering whether the issues causing removal are possibly correlated with some aspect of program participation and/or savings. This issue could lead to biased results. If removal is greater than 5%, then the analysis should include a table that compares the analysis group to the program participant population on available data (such as house characteristics, program measures, and pre-retrofit usage).

4.5.2.2 Weather Data

Weather data are used in the consumption data analysis in two ways:

- In models that relate consumption to weather, the observed weather data are matched to the meter read intervals to provide predictor variables.
- The model estimated with actual weather is calculated at normal-year weather levels to provide usage and savings in a normal or typical year.[20]

Use either primary National Oceanic and Atmospheric Administration (NOAA) or weather stations managed by the utility (and trusted by utility analysts) as the source for weather data. Some utilities maintain weather series (both actual and normal/TMY) for internal use, and it is generally best to use a utility's weather resources to produce evaluation results that are consistent with other studies within the utility. Many utilities are choosing to use norms constructed from fewer than 30 years, as are the standard NOAA norms.

A consumption data analysis requires both actual and normal (or TMY) weather data from a location near each premise. The actual weather data must match the time interval of each meter reading interval. Both actual and normal/TMY weather data used for each site should come from the same weather station. Only annual TMY degree days are required for annual analysis results. This protocol recommends calculating the annual monthly normal degree days for the purpose of plotting model fit values.

4.5.2.3 Weather Data Preparation

Depending on the source, weather data may need additional preparation. Limited missing data can be filled by the simple interpolation. If the amount of missing data is sufficient to trigger

[20]NOAA produces 30-year normal weather series composed of average temperature for each hour over the time period. These norms are updated every decade. NREL produces TMY data series. These data are not average values but a combination of typical months from years during the time period. The TMY data also cover a shorter time period.

concern regarding a weather data source, consider using a more distant but more complete weather station as an alternative.

Create a graph to identify anomalies, gaps, and likely data errors. Weather data issues tend to be obvious visually. Missing data and technical failures look very different than naturally random weather patterns. For each weather station used in the analysis, plot the following information over the analysis time span: minimum, maximum, and average temperature versus day of year. If multiple weather stations are used across a large region, plot the different stations on a single graph.

4.5.2.4 Tracking Data

The program tracking data provide the participant population, the installation date or a proxy such as paid date, and the number and type of measures for which savings are claimed. Frequently, the original consumption data request is made based on the population defined by the tracking data. Additional information in the tracking database may serve as a resource for other elements of the analysis:

- If a variety of measures were installed and there is a sufficient mix of different combinations of measures, it may be possible to develop savings estimates for some individual measures. In this situation, focus the evaluation on the measures with greater expected savings for separate estimates of savings.
- The date of a measure's installation both provides the date at which the change in consumption took place ***and*** identifies the billing interval(s) that will be blacked out. The tracking database, however, may contain the installation confirmation date, the date of payment, or some other date loosely associated with the time at which consumption actually changed (rather than the explicit installation date). The evaluator should consult with the program staff to determine what the different recorded dates refer to and when actual installation could have occurred in relation to these dates.

  Also, it may be necessary to black out multiple billing periods. Multiple installation dates at the same site may require a longer blackout period or may make the site untenable for simple pre-post analysis. If the blackout period does not encompass the dates of all program-related changes to consumption, then the pre-post difference will be downwardly biased.
- The tracking data may also be a useful resource regarding the characteristics of participant homes. Frequently, program databases capture home square footage, number of floors, existing measure capacity, and efficiency. These data are primarily useful in the pooled approach if they are only available for current participants.
- Tracking data from previous years may be used to define a control group for a two-stage analysis.

4.5.3 Analysis Dataset

Using the account numbers in the two datasets, the final analysis dataset combines the tracking data and the consumption data with the weather data. Weather data are attached to each consumption interval, based on the days in a read interval. The combined data have a sum of the daily degree-days for each unique read interval, based on start date and duration. If the variable degree-day base approach is used, this process must be repeated over the range of heating and cooling degree-day bases. To produce average daily consumption and degree days for that read interval, the read interval consumption and degree-day values are divided by the number of days in the interval.

Because of the complication of matching weather to all the unique read intervals, some evaluators resort to calendarized data.[21] Except in special cases, calendarization should not be used for this kind of analysis because it undermines the direct matching between consumption and degree days that is the basis of consumption data analysis. Multiple meter and multifamily analyses are examples of situations where calendarization may be the only way to aggregate data series on different schedules.

4.5.3.1 Analysis Data Preparation

A number of additional data preparation steps are required when the three data sources (tracking, billing, and weather) have been combined. These limit the analysis data to only the data to be included in the model.

- ***Participant Data Only.*** Confirm that the consumption data in the analysis dataset is only for the household occupant who participated (or will participate) in the program.
- ***Blackout Interval.*** Remove from the regression the full read interval within which the installation occurred. If the installation timing is not explicitly indicated in the tracking system—or if installation occurred in stages over several weeks or had ramp-up or ramp-down effects—it may be necessary to extend the blackout interval beyond a single read interval.
    - For a single, relatively simple measure (such as a furnace), a single blackout month is sufficient.
    - For more complex installations (longer-term single installations or multiple installations), a multiple-month blackout may be more appropriate.

  The change in consumption will be biased in a downward direction if part of the transition interval is included as either pre- or post-installation typical consumption. In most instances, the only negative aspect of increasing the blackout interval is the corresponding decrease in either pre- or post-installation readings.
- ***Sufficient Data for a Site.*** Count the number of data points in the pre- and post-blackout periods for each individual site consumption data series. To create a view of the classic seasonal consumption data patterns, plot a representative sample of daily average consumption data by read date. Daily average consumption plotted by temperature replicates the underlying structure of the consumption data analysis. Plotting the estimated and actual monthly values in both formats is the most effective way to identify unexpected issues in the data and to reveal issues related to model fit.

  Ideally, a full year of consumption data is available for each site for the pre- and post-blackout periods.

[21] Calendar month consumption is estimated as a weighted average of the bill readings that cover that month.

- For individual site analysis of electric consumption, a minimum of nine observations spanning summer (July and August), winter (January and February), and shoulder seasons are recommended for each site in each time period (pre- and post-installation). For gas consumption, six observations spanning at least half of a winter and some summer are the minimum.
- For a pooled analysis, sites with fewer observations or fewer seasons represented can be included (a minimum of six in each period). However, it is important to have all seasons represented in both time periods and across all premises in the pooled model.
- Bimonthly data provide a particular challenge for consumption data analysis. In a year of data, all seasons are represented, but the number of data points is halved. For analysis of gas consumptions, a minimum of one year each of pre- and post-installation data is essential. For analysis of electric consumption, two years each of pre- and post-blackout data are better.

5 Looking Forward

As discussed in Section 2.1, more granular AMI data are increasingly available to evaluators pursuing consumption data analysis. These data bring new opportunities and new challenges to evaluation. While the more granular data offer the possibility of estimating the peak period kW effects and time-differentiated energy efficiency impacts (kWh) for a program, they also increase the breadth and complexity of modeling approaches and the computing power required to produce results. Also, although the granularity of available consumption data is increasing, the other data available for inclusion in a typical evaluation model—tracking data, weather, etc.—remain mostly the same.

Whole building evaluation will benefit substantially by incorporating the learning from site-level modeling efforts that have been pursued for years in the commercial sector where interval data have been available, as well as from demand response/direct load control modeling efforts that have used both end-use and whole-building data for the purpose of modeling short term load curtailments. A protocol addressing the use of AMI data for consumption data analysis will contend with almost all the issues put forward in this chapter as well as the additional challenges revealed with the more granular data—the diurnal patterns combined with the unique thermal dynamics of each building.

As consensus is reached on the best practices in the use of AMI data for consumption data analysis, an additional chapter, or a substantially expanded version of this chapter, will be needed to capture these practices.

6 References

Fels, M.F., ed. (February/May 1986). Energy and Buildings: Special Issue Devoted to Measuring Energy Savings: The Scorekeeping Approach. (9:1&2).

Fels, M.F.; Kissock, K; Marean, M.A.; and Reynolds, C. (January 1995). PRISM (Advance Version 1.0) Users' Guide. Center for Energy and Environment Studies. Princeton, New Jersey.

Goldberg, M. L., Fowlie, M.; Train, L.; and Agnew, G.K. (2017, in revision). A White Paper: Mitigating Self-Selection Bias in Billing Analysis for Impact Evaluation. DNV GL, submitted to PG&E.

Goldberg, M. L., Fowlie, M.; Train, K.; and Agnew, G.K. (2017). Not Just Another Pretty Formula: Practical Methods for Mitigating Self-Selection Bias in Billing Analysis Regressions. Proceedings of the International Energy Program Evaluation Conference. Baltimore, Maryland.

Jacobson, D. (2017). *Chapter 5: Residential Furnaces and Boilers Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68561. http://www.nrel.gov/docs/fy17osti/68561.pdf

Randazzo, K.; Ridge, R.; and Wayland, S.(2017, in revision). Observations on Chapter 8 of the Uniform Methods Project: A Discussion of Comparison Groups for Net and Gross Impacts. Opinion Dynamics, submitted to PG&E.

Stewart, J. and Todd, A. (2017). *Chapter 17: Residential Behavior Protocol. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68573. http://www.nrel.gov/docs/fy17osti/68573.pdf

Violette, D. M.; Rathbun, P. (2017). *Chapter 21: Estimating Net Savings – Common Practices, The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68578. http://www.nrel.gov/docs/fy17osti/68578.pdf

7 Resources[22]

ASHRAE. (TBD). *Guideline 14-2002R*. (Revision of *Guideline 14*, currently in process).

ASHRAE. (2010). *Performance Measurement Protocols for Commercial Buildings*.

ASHRAE. (2002). *Guideline 14-2002 Measurement of Energy and Demand Savings*.

ASHRAE. (2004). *Development of a Toolkit for Calculating Linear, Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models*. ASHRAE Research Project 1050.

[22] Some resources recommended by ASHRAE.

Chapter 9: Metering Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

Dan Mort
ADM Associates, Inc.
Sacramento, California

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68565
September 2017

Contract No. DE-AC36-08GO28308

Chapter 9: Metering Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

Dan Mort
ADM Associates, Inc.
Sacramento, California

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68565
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

These methods, processes, or best practices ("Practices") are provided by the National Renewable Energy Laboratory ("NREL"), which is operated by the Alliance for Sustainable Energy LLC ("Alliance") for the U.S. Department of Energy (the "DOE").

It is recognized that disclosure of these Practices is provided under the following conditions and warnings: (1) these Practices have been prepared for reference purposes only; (2) these Practices consist of or are based on estimates or assumptions made on a best-efforts basis, based upon present expectations; and (3) these Practices were prepared with existing information and are subject to change without notice.

The user understands that DOE/NREL/ALLIANCE are not obligated to provide the user with any support, consulting, training or assistance of any kind with regard to the use of the Practices or to provide the user with any updates, revisions or new versions thereof. DOE, NREL, and ALLIANCE do not guarantee or endorse any results generated by use of the Practices, and user is entirely responsible for the results and any reliance on the results or the Practices in general.

USER AGREES TO INDEMNIFY DOE/NREL/ALLIANCE AND ITS SUBSIDIARIES, AFFILIATES, OFFICERS, AGENTS, AND EMPLOYEES AGAINST ANY CLAIM OR DEMAND, INCLUDING REASONABLE ATTORNEYS' FEES, RELATED TO USER'S USE OF THE PRACTICES. THE PRACTICES ARE PROVIDED BY DOE/NREL/ALLIANCE "AS IS," AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL DOE/NREL/ALLIANCE BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER, INCLUDING BUT NOT LIMITED TO CLAIMS ASSOCIATED WITH THE LOSS OF PROFITS, THAT MAY RESULT FROM AN ACTION IN CONTRACT, NEGLIGENCE OR OTHER TORTIOUS CLAIM THAT ARISES OUT OF OR IN CONNECTION WITH THE ACCESS, USE OR PERFORMANCE OF THE PRACTICES.

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Suggested Citation

Mort, D. (2017). *Chapter 9: Metering Cross-Cutting Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68565. http://www.nrel.gov/docs/fy17osti/68565.pdf

Acronyms

CT	current transformers
DOE	U.S. Department of Energy
EEM	energy efficiency measures
EMS	energy management system
HVAC	heating, ventilating, and air conditioning
IPMVP	International Performance Measurement and Verification Protocol
IR	infrared
IWC	inches of water column
kW	kilowatt
kWh	kilowatt-hour
M&V	measurement and verification
MCC	motor control center
NIST	National Institute of Standards
PF	power factor
psi	pounds per square inch
psig	pounds-per-square-inch gauge
RH	relative humidity
RMS	root mean square
RTD	resistive temperature devices
UMP	Uniform Methods Project
V	voltage
Vac	alternating current voltage
Vdc	direct current voltage
VFD	variable frequency drive
VSD	variable speed drive

Table of Contents

List of Figures

List of Tables

1 Introduction

Metering is defined as the use of instrumentation to measure and record physical parameters. In the context of energy-efficiency evaluations, the purpose of metering is to accurately collect the data required to estimate the savings attributable to the implementation of energy efficiency measures (EEMs).

Estimated energy savings are calculated as the difference between the energy use during the baseline period and the energy use during the post-installation period of the EEM. This chapter describes the physical properties measured in the process of evaluating EEMs and the specific metering methods for several types of measurements. Skill-level requirements and other operating considerations are discussed, including where, when, and how often measurements should be made. The subsequent section identifies metering equipment types and their respective measurement accuracies. This is followed by sections containing suggestions regarding proper data handling procedures and the categorization and definition of several load types. The chapter concludes with a breakdown of recommended metering approaches by load category, which is summarized in Tables 2 through 7.[1]

[1] As discussed in the section "Considering Resource Constraints" of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

2 Metering Application and Considerations

Metering allows for the quantification of the energy use of a load. Metering can also record parameters—such as hours of operation, flows, and temperatures—used in the calculation of the estimated energy savings for specific end uses. (The recording of such parameters through metering methods is also referred to as “monitoring.”)

2.1 Identifying Scope

To optimize equipment and labor costs, it is important both to identify the scope of a metering procedure and to measure the key parameters required for estimating energy usage and savings. Although it may be possible to measure numerous parameters in a given facility, a metering procedure should focus on those parameters required for energy savings estimations. Therefore, to identify the necessary loads or parameters for the calculation, the savings estimation methodology for the EEM should be developed before the installation of metering instruments. If the data are a critical aspect of the estimated savings calculation, a redundant measurement or an additional proxy measurement for the parameter of interest may be considered. However, such considerations should be made within the context of ensuring a practical and cost-effective metering process.

The specific metering equipment for the job should be selected before visiting the site to install the meters. When installing more than one piece of equipment as part of an EEM, refer to Chapter 11: *Sample Design Protocol* to determine how many units need to be metered.

2.2 Ensuring Precision and Verification

The accuracy of a measurement is typically proportional to the cost of the instrument and the installation method. Additionally, such factors as measurement location, monitoring duration, and sampling interval also impact the accuracy of the results. For a given measurement or parameter, the necessary precision is an important consideration in the savings estimation. Higher-cost metering equipment may be required, depending on site and project characteristics. Further explanations regarding savings estimation analyses are detailed in other chapters.

Verification of the collected data is an essential aspect of ensuring an accurate metering process. Key best practices for data verification are these:

- Review the data to: (1) verify that they are complete and correct, and (2) identify readings that appear inappropriate or notably atypical for the specific system.
- If the readings appear to be incorrect, conduct cross-checks with other sensors or meters. Additionally, review the assumptions that were made when planning the metering to assess their validity and appropriateness.
- If the cross-checks do not validate the data, calibrate the equipment to match other metering instruments. Alternatively, determine whether the sensor or meter needs to be replaced.
- Validate the metering equipment results with facility-installed instruments, as needed, as another method of cross-checking. If the facility has data recording capability or an energy management system (EMS), readings from those systems can be used for

reference. Ultimately, however, these measurements must be objectively validated against independent metering equipment.

- Assign the data-collection responsibilities to a specific individual who will determine the design and structure of the metering process.
- Review the retrieved data for completeness and accuracy before incorporating it into the final analysis.

Before installing a meter, test it to ensure it is working properly and making the intended measurement. Use this checklist as a guide:

1. If meter operates on batteries, are the batteries in good condition, and do you have a backup set? Is the meter properly powered?
2. Is the meter clock synchronized to National Institute of Standards (NIST)[2] and local time zones?
3. Are all the settings on the meter correct?
4. Are sensors properly attached and in place?
5. If possible, did you turn the meter load on and off after installation and before removal to obtain a signal that the meter is capturing the correct equipment?

[2] www.nist.time.gov

3 Type of Measurement

Measurement types can be categorized by the associated physical properties they represent. Individuals conducting measurements should understand the purpose of the measurement. This section describes these properties and their respective measuring methodologies. The corresponding equipment descriptions are included in a subsequent section.

3.1 Electrical

Electric power and energy are typically the most important measurements for savings evaluations. As electric power is commonly a direct measurement of the energy use of a load, it may be the only measurement needed to determine savings between a base case and high efficiency measure.[3]

The common unit of power is kilowatts (kW). The common unit of energy is kilowatt-hour (kWh). Energy is power used during a unit of time. Other electrical measurements are voltage (V), current in amperes (A)[4], and power factor (PF). Although direct current voltage (Vdc) is used to power some types of equipment, utility transmission to customers occurs in the form of alternating current voltage (Vac). For this discussion, A and V are expressed in terms of alternating current, and the values measured or recorded are the root mean square (RMS) values. In general terms, RMS is the common presentation of alternating current electrical measurements. Apparent power (V·A) multiplied by the power factor equals the true power (W=V·A·PF). Power factor is given by the following:

- For perfect sinusoidal waveforms, the power factor is the cosine of the angle of the phase shift between the current and the voltage.
- If the voltage and current waveform are non-sinusoidal, the definition of power factor is (V·A)/W.

3.1.1 Considerations

There are important safety and metering considerations associated with conducting power measurements. Only an electrician, an electrical engineer, or a technician with training and proper equipment should be allowed to work in live electrical panels. Also, the individuals conducting this work should know and follow codes and guidelines provided by the National Electric Code (NEC), the Occupational Safety and Health Administration (OSHA), and the National Institute for Occupational Safety and Health (NIOSH). Additionally, personal protective equipment (PPE) that complies with National Fire Protection Association (NFPA) 70E should be worn to protect against arc flash in open electrical cabinets.

Electrical measurements should be limited to 600 V or less. Due to spark gaps from the high voltage, only electrical linemen with special training and equipment should work on systems above 600 V. Some facilities have existing current and voltage sensors in place on systems greater than 600 V that can be safely utilized to make measurements.

---

[3] Note that power metering is also referred to as kW metering.
[4] Current metering is also referred to as Amp metering.

Current metering rather than power metering can be considered if:

- The load has a stable or well-defined power factor and the interval of recording is short relative to the system cycle
- The metering is only to determine operating hours.

When conducting current metering, additional analysis is needed to convert current data to power data.

Harmonics are produced by electronic loads. These non-sinusoidal waveforms can only be accurately measured by meters designed to make true RMS measurements.

3.1.2 Single Phase vs. Three-Phase Loads

The two common standard voltages utilities provide to most commercial customers are three-phase 120/208 V or 277/480 V. The term "277/480 V" signifies that the voltage from any one of the phases to ground is 277 V and the voltage from one phase to another phase is 480V.

- The two main types of three-phase electrical systems are wye and delta.
- Wye systems are three-phase and four-wire, where the fourth wire is neutral.
- Delta systems are three-phase and three-wire.

There are several less common variations with grounding differences relative to the active voltage legs.

Residential supply voltage is 120/240 V and is single phase. It uses a three-wire configuration consisting of two hot legs and one neutral.

While lighting is a single-phase load, most motors are three-phase loads. Three-phase motors are assumed to be balanced, which means the current draw is equal in each of the three phases. In practice, however, the three-phase currents are not always identical.

3.1.3 One Time Power Measurements

Power measurements require the opening of electrical panels to gain access to where the insulated conductors or wires make electrical contact with safety devices such as breakers or fuses. When conducting power measurements, the technician or engineer should reference the connection diagram provided by the meter manufacturer for the specific supply voltage.

Power measurements also require the simultaneous detection of both current and voltage. This is typically achieved by placing a clamp-on current probe around the conductor of a given phase. After placing one of the meter voltage leads in contact with an exposed junction of the same phase, connect the other lead to neutral or ground.

For handheld meters that can only make measurements on one phase at a time, measure each phase separately. For three-phase systems without a stable ground—or in situations where there are doubts about the configuration—make measurements with a portable three-phase power meter. The total power of the system is defined as the sum of the power for all three phases.

When conducting power measurements, document the V, A, W, and PF measurements for each phase. For loads where current metering is sufficient, metering one phase and conducting one-time measurements on all three phases is required. To determine power from current metering, the load must have a power factor that is stable or a well-defined profile with loading. Taking one-time measurements that include power factor at multiple load conditions (varying current) improve the power analysis.

3.2 Temperature

Temperature is an indirect parameter that is incorporated into the calculation of energy use or estimated savings for some types of EEMs. Temperature sensors can be designed to measure gases, liquids, or solids. Typical applications for temperature measurement include ambient air, supply or return air, air or other gas in an enclosed space (such as near a thermostat), combustion gas, supply and return of fluids (such as chilled water), water heaters or boilers, steam condensate, and refrigerant lines.

Unless otherwise specified, “air temperature measurement” always refers to a dry-bulb temperature measurement. Wet-bulb temperature is defined as the temperature of a wet surface when water is evaporated from that surface for a given condition. This temperature is always lower than dry-bulb temperature, unless the air is completely saturated with water vapor. In this case, the two values would be equal. Dry-bulb and wet-bulb temperature are used together to determine the humidity or moisture in the air. Humidity is used in energy-use estimations for various air-conditioning systems.

3.2.1 Considerations

There are no specific qualifications required for the personnel who conduct temperature measurements, but these individuals should understand the purpose of the measurement.

When making temperature measurements, consider such factors as these:

- Weather conditions
- Location, sunlight exposure
- Heat radiating from nearby hot surfaces
- Contact with the media being metered
- Insulation from ambient conditions
- Air movement stagnation or stratification.

3.2.2 Outdoor Air Temperature

Outdoor temperature measurements are notably vulnerable to the surrounding environment, so this effort requires these additional precautions:

- Protect the temperature sensor from moisture, such as blowing rain.
- Use a radiant shield to protect the sensor from direct sunlight and reflected surfaces.
- Place the sensor in a well-ventilated location so that neither air stagnation nor stratification contributes to the temperature measurement.

3.2.3 Duct Air Temperature

Temperature sensors in ducts should be placed where the air is well mixed. For example, the supply air temperature should not be immediately downstream from the evaporator coil; instead, it should be several duct diameters downstream. To determine the best sensor location, take spot measurements in a traverse. This can be a challenge in large ducts when deploying averaging sensors. (An averaging sensor is composed of an array of individual sensors that can be placed as a web or matrix of points in a duct cross-section to measure the average temperature in the space.)

3.2.4 Liquid Temperature

Water (or glycol) temperature in pipes can be measured by: (1) inserting temperature probes into the liquid, (2) placing probes in thermal wells, or (3) placing probes on the pipe surface. Both the physical configuration of the existing piping and the willingness of the customer or contractor to drill into pipes typically dictate the appropriate installation method. The costs are relatively comparable for each approach.

- **Insertion probes** make direct contact with the liquid and, thus, provide the most accurate measurement. However, insertion probes can be problematic, because they require either (1) an unused tap on the pipe with a port that has a self-sealing pressure gasket (Pete's Plug) where the probe can be inserted or (2) the installation of a costly hot tap on the pipe (a technique that allows insertion of a probe into a pressurized pipe without having to shut down the system).

- **Thermal wells** are an effective alternative to insertion probes. Some pipes have pre-existing thermal wells strategically placed to measure supply and return temperature; however, these wells are often already in use by system or process controls. If a thermal well is available, apply thermal grease to the probe to increase overall conductance.

- **Surface mount probes** mounted on a pipe—for pipes that are not plastic—are an alternative to thermal wells. Apply thermal grease between the probe and the pipe surface (on the underside of a horizontal pipe) to eliminate any air gaps. Then, use a minimum of one inch of insulation over the probe so that the probe registers the temperature of the pipe contents rather than the air.

- **Infrared (IR) thermometers** can be used to make instantaneous measurements of surface temperatures. Although the laser pointer on an IR thermometer produces only a small red dot, the surface area being measured is significantly larger. For example, if the distance-to-target ratio for the meter is 12:1, then at a distance of three feet, the surface area of measurement is three inches in diameter.

3.3 Humidity

The common unit of humidity is the percentage of relative humidity (%RH). Relative humidity is a measure of the relative amount of water vapor in the air for a given condition, versus the capacity of the air to hold water vapor at that same condition.

Humidity is measured when estimating the enthalpy or energy content of air in a heating, ventilating, and air-conditioning (HVAC) system. Humidity is also measured to determine comfort conditions using psychrometric charts. Outdoor humidity can be used to provide a measurement of ambient conditions. The placement requirements for humidity sensors are the same as those for ambient air temperature sensors. It is important to use measurements from steady-state conditions when using humidity sensors, because these sensors have a slow response time.

3.4 Flow of Liquids and Gases

The common unit of flow for liquids is gallons per minute (gpm), and the common unit of flow for gases is cubic feet per minute (cfm).

3.4.1 Water Flow

Measuring the flow rate of water or glycol in a chilled water loop is one parameter in determining the output of a chiller. Typically, a mechanical contractor is needed to install a water flow meter. A flow meter should be installed on a straight uniform section of pipe at least 15 diameters long, with the meter 10 diameters downstream from the last bend or transition, so as to minimize turbulence in the liquid stream.

A passive measurement of fluid flow can be made using an ultrasonic flow meter at a point where there is no pipe insulation. Ultrasonic flow meters, which are applied to the outside of the pipe, send pulsed sound signals through the fluid. These signals measure the flow of water-based liquids in pipes without interrupting the flow (as a flow sensor inside the pipe would). Note that ultrasonic flow meters are typically very costly and require experience to use, which should be considered when designing the metering process.

An alternative to water flow measurement entails measuring the pump motor electric demand to determine motor loading. The electric demand and another variable (such as pressure) are then cross-referenced with the manufacturer's pump curve data to calculate flow rate. While this option is a lower-cost solution, the resulting measurement is generally not as accurate using a water flow meter.

3.4.2 Duct Airflow

Airflow measurements are most commonly needed for ducts carrying conditioned air, and these measurements can be made by anyone trained in the technique. Note that gas or airflow rates should be normalized to standard temperature and pressure conditions (68°F and 14.7 psi).

The preferred methods for measuring airflow rate use these technologies. In residential applications, the first three of these options are viable; however, for commercial duct systems, the fourth option may be the only viable choice.

- A calibrated adjustable-speed fan at the return register
- A pitot tube array at the air filter
- A matrix of transverse air velocity measurement points in a long straight cross-section of the duct
- A flow capture hood at the return or supply registers (a less reliable technique).

The matrix of air velocity measurements is more costly, due to labor and preparation time. For this approach, select a straight uniform section of duct at least 15 diameters long, with velocity measurements that are made 10 diameters downstream from the last bend or transition, so as to minimize turbulence in the air stream.

Airflow in a compressed air system can be measured with a mass flow rate sensor, which compensates for density with respect to pressure. The sensor should be installed only when the system has been shut down by an individual having the appropriate mechanical experience.

3.4.3 Natural Gas

Natural gas can be measured by installing a utility-style meter on the gas-fired equipment. Generally, there are few opportunities to meter this equipment, however, because of the cost, difficulties in coordination of installation with the proper licensed trades, safety considerations (including clearing pipes of all residue gas before installation), and limited installation accessibility. In some cases, existing utility meters that supply gas to only the measure in question can produce a pulse for recording.

Natural gas-fired equipment that has a constant burner flow rate can be measured using the fine resolution dial on the utility meter and a stop watch ***if*** all other gas appliances are off during the test. Note that equipment gas lines should be turned off during the installation, and a qualified gas fitter should conduct the installation.

3.5 Pressure

The common unit of pressure is pounds per square inch (psi). Although pressure is not used to estimate energy use directly, it can be incorporated as a normalizing measurement or used to calculate the efficiency of fans or pumps. An example of this is measuring the pressure in a compressed air system before and after a variable frequency drive is installed.

3.5.1 High Pressure

High pressures occur in fixed volumes such as tanks, refrigerant loops, and pumping systems. Instances where high-pressure measurement is required include compressed air equipment, water pumping stations, and refrigerant lines. Place high-pressure sensors on a port with a valve so they can be installed without shutting down the system. A qualified mechanical contractor should conduct the installation of the port.

3.5.2 Low Pressure

Low-pressure air pressure measurements encompass static, dynamic, and barometric. Static and dynamic pressure measurements can be taken in air ducts to gauge airflow rates. These low-pressure measurements occur where the air is not enclosed in fixed volumes.

Static pressure measurement in a combustion ventilation pipe is used to determine whether adequate draft is available to exhaust combustion byproducts.

A technician can install a static pressure gauge in a duct system to measure static pressure change across the fan.

3.6 Light

Light level (or illuminance) is commonly measured in units of either foot-candles (fc) or lux. While illuminance is not used to estimate energy savings directly, it is often used to verify that the pre- and post-lighting equipment either supply an equivalent amount of light or meet certain end-use requirements. However, if, after the EEM is installed, there is a decrease in light levels to below code or recommended levels, illuminance measurements can be used to justify a reduction in final savings. Conversely, if light levels increase above code or recommended levels after an EEM is installed, the illuminance measurements justify applying additional savings. There are no specific qualifications required for personnel conducting illuminance measurements.

3.6.1 Considerations

When making illuminance measurements, consider both the working conditions and background daylight conditions. Take measurements at the level of the working surface, usually a desk or table. Also, account for ambient light or daylight by taking measurements when the EEM lighting is on and again when it is off. The difference in the two values is the illuminance attributable to the EEM lighting.

3.7 Status or Event

Some measurements are in the form of bi-level logic that identifies whether (1) a load is on or off or (2) a switch or door is open or closed. These are cost-effective approaches to metering a piece of equipment's time-of-use hours of operation. So long as these loggers are not placed in live electrical panels, there are typically no specific qualifications required for personnel placing status loggers; however, training is recommended.

Analyzing on/off status records of a load (such as lighting or motors) is a convenient method of measuring hours of operation. A valve or damper position may also be needed to determine operating mode of an HVAC system.

3.8 Normalizing Conditions

In many cases, to normalize the energy use of the EEM, it is necessary to collect additional data. Energy use for both a baseline and a post-implementation period should be normalized if any specific conditions differ between the two periods. For weather-dependent loads, typical meteorological year weather data are used to normalize the energy savings.

Normalizing data can either be measured and recorded from the equipment itself or collected from facility management, if necessary. Normalizing parameters typically include:

- Production volume
- Processed weight
- Sales
- Occupancy
- Set points
- Ambient temperature
- Weather
- Flow
- Pressure
- Speed
- Frequency
- Alternative operating modes

4 Levels of Measurement

Electric loads should be metered at the level appropriate for the type of EEM. The levels may be defined through aggregations of:

- Like loads (such as lighting)
- Measurements of electric load in an area (whole panels) that is a subset of the utility meter
- Measurements of a system (such as pumps, fans, and compressors of an HVAC system)
- The utility meter itself.

4.1 Single Loads

Measurements on single loads—such as motors—are performed on the conductors serving the unit exclusively. The electrical measurement can be made in (1) the motor control center (MCC) panel serving the load, (2) the disconnect box at the motor, or (3) the variable speed drive (VSD), if applicable.

In the case of a VSD, the measurement should not be made on the conductors between the VSD and the motor. Metering inside a VSD can be problematic in that the drive can cause interfering signals in metering equipment even if the metering is upstream of the drive. For this reason the preferred location to meter VSDs is at the MCC.

4.2 Aggregation of Like Loads

Lighting is generally updated as a retrofit throughout a wide area of a facility or throughout an entire facility, so metering a representative sample rather than conducting metering for a census of fixtures usually suffices.

When selecting a metering sample for an end use, the sample should be categorized by operating hours or by the variation in load. For example, lighting within a facility should be stratified by area types with different operating schedules or patterns. After the number of fixtures in each specific area type has been determined, the sample size can be quantified. (See Chapter 11: *Sample Design Protocol*.)

Measuring electric loads by area or by whole-panel metering is useful when developing an hourly use profile. Specifically:

- Meter whole electric panels that exclusively serve end uses of interest.
- For panels that also serve other end uses, account for those end uses by metering the panels and subtracting that load from the total, or by other means (such as engineering estimates).

When using building energy simulation models, area metering is useful for determining internal load profiles for inputs.

4.3 Measurements of a System

If the end use is a chiller, take measurements related to the operation of the chiller. The system may contain the chiller, chilled water and condensate pumps, cooling tower fans, and air handlers. These measurements may include power input and thermal output—as measured by supply—and return chilled water loop temperature and water flow rate. Note, however, that chilled water loop measurements may be hampered by pipe insulation. Conversely, condenser water pipes may not have insulation and, thus, they may provide greater accessibility for surface mounted temperature probes and externally mounted ultrasonic flow meters.

5 Duration of Measurement and Recording Interval

Measurement duration is classified into three categories: instantaneous, short-term, and long-term. Each duration category has a purpose and should be selected based on the specifics of the EEM and magnitude of the load.

5.1 Instantaneous

Instantaneous measurements (also known as “spot measurements” or “one-time measurements”) are used to (1) quantify a parameter that is expected to remain constant or (2) calibrate instruments that will collect data over a period of time. These measurements are generally made using handheld instruments at the location of the parameter of interest; however, they can also be made using instruments installed as part of a system.

5.2 Short-Term

Short-term measurements are conducted to record the variation of a parameter over a period of time. To capture at least two cycles of the load or parameter of interest, instruments performing short-term metering are put into position for periods ranging from several hours to one month.

For example, although the lights in a business operation may turn on and off from day to night, the overall lighting in most business operations has a weekly cycle, because the weekend schedule generally differs from that of weekdays. Typically, a two-week period of data is collected, so that data from the second week can confirm the pattern of the first week. However, if the loads vary during the year, then long-term metering periods should be considered. Also, the appropriate monitoring period should be selected to include peak loads if demand savings estimates are part of the measurement and verification (M&V) effort. Cooling loads, for example, should be monitored during the hottest part of the year.

5.3 Long-Term

Long-term measurements are conducted to record variations of a parameter that occur over a period generally ranging from one month to one year. Instruments performing long-term metering are typically installed at sites that are:

- Weather-dependent (such as HVAC loads)
- Seasonal (such as agricultural processing)
- Operate on planned schedules (such as educational facilities).

5.4 Recording Interval

“Measurement time resolution” refers to the length of intervals used during data collection. Recording intervals are at one or more minutes (often in increments of 5, 10, or 15 minutes), although many loggers allow other time intervals.

Use intervals that are integer divisors of 60 to facilitate processing the data into hourly totals. Also, some equipment types average or sum the values for the interval, while other types only record an instantaneous reading at the end of each interval. Instantaneous readings at the end of each interval should only be used if the measured parameter is changing slowly with respect to the interval duration or if enough interval points are captured to provide statistical significance.

For most load types, 15-minute aggregate interval data provide sufficient time resolution to capture reaction of the load to the controlling conditions. (Note that utility electric meters are also designed to record peak kW on 15-minute intervals.) Where recorder memory capacity allows shorter intervals, it is possible to capture profiles of loads with short cycle times. For loggers that only provide instantaneous readings, the interval length should be short enough to capture at least five recordings per cycle of the load. For example, if an air-conditioning unit cycles once every 25 minutes, then the recording interval should be five minutes or less.

As technology advances and measurement equipment increasingly contains more memory storage, it is possible to collect data in very small time intervals. However, additional data are not likely to increase the accuracy of the savings estimation significantly, and there is typically an increase in the costs associated with analysis processing time.

For loggers that record both the date and time stamps of events, the time uncertainty is a combination of the reaction time of the sensor and the time stamp resolution.[5] Logger clock drift is generally small but should be checked at the time meters are retrieved in order to document any drift during the data recording period.

[5] Time stamp resolution is generally one second.

6 Equipment Types

This section, which discusses various metering devices, is categorized by the parameter the devices are used to measure. (Note that the terms "recorder" and "logger" are often used interchangeably to describe metering equipment.)

There are two main categories under which metering equipment are typically classified:

- Type of measurement (This equipment can be sub-categorized by dedicated single measurement or by multi-purpose and multi-channel.)
- Metering function, such as sensor-only, instantaneous readout meter, recording meter, or recorder only.

Instrument accuracy is typically not represented by a single value. In most cases, accuracy is provided as a plus or minus (±) percentage of the reading and is only appropriate for a prescribed range of values from the full-scale (fs) reading. Also, the accuracy may be different for various ranges.

Most meters use proprietary software to set proper data collection parameters, recording intervals, clock settings, etc. The manufacturer's software must also be used to retrieve data from the meter and then export it to other usable formats (i.e., text or spreadsheets).

Follow local codes when metering with any type of equipment. This is not only for the safety of the technician but also for the safety of others where equipment is located.

6.1 Electrical

Electrical measurement equipment can be categorized as:

- Handheld (or portable) power meters
- Watt-hour transducers
- Meter recorders
- Current transformers (CT).

6.1.1 Handheld Power Meters

Select handheld power meters measure true RMS volts, amps, watts, and power factor. Ideally, these meters have a digital display of at least 3.5 digits and measure power to an accuracy of ±2.5% or better. The voltage, current, and power factor accuracy will all be greater than this because the combination of the individual measurement accuracies is used in determining the power accuracy.

A clamp-on current sensor can either be an integral part of the meter or a separate sensor connected to the meter with a wire cable. The jaws of the current sensor should be able to hold all of the conductors on the phase of the load being measured.

6.1.2 Watt-Hour Transducers

Watt-hour transducers only measure the power or energy use, so they need a separate logger to record the use over short-term or long-term metering. Watt-hour transducers typically produce a pulse output in which each pulse represents a predetermined number of kWh, depending on the system voltage and CT ratings. Following the recording, a multiplier is applied to scale the pulse output into units of kWh. (Review manufacturer specifications to determine the multiplier.)

The watt-hour transducer should have an accuracy of ±0.5% or better. Note that the CT accuracy must be added to the transducer accuracy to determine the power measurement accuracy. In the event that the two pieces of equipment are correlated, the accuracies are added together. If they are not correlated, then the combined accuracy is the square root of the sum of the squares of the individual accuracies.

Current sensors are typically selected separately and are sized based on the peak current the load will achieve during the metering. The signal output types for watt-hour transducers include 4-20mA, 0-10Vdc, LonWorks, Modbus, and BACnet. Some of these are more appropriate for EMS than for short-term monitoring.

6.1.3 Meter Recorders

Meter recorders both measure and record on the same instrument. The meter selected should measure true RMS power. Current sensors are generally selected separately and are sized based on the peak current that the load will achieve during the metering. To determine the accuracy of the power measurement, combine the CT accuracy with the transducer accuracy. (As mentioned in the previous section, the watt-hour transducer should have an accuracy of ±0.5% or better.)

In general, meters and sensors must be fully contained in the electrical panel; however, if the voltage exceeds 50 V, the meters and sensors will require wiring to be placed inside of conduit to the meter. Cables conducting low-voltage sensor signals (such as pulse outputs, 333mV CT leads, or communication signals) do not need to be inside of conduit. Follow the manufacturer's directions for connecting CTs and voltage leads, as these instructions differ, depending on number of phases and wires, voltage, and configurations (such as wye, delta, and high-leg delta).

6.1.4 Current Transformers/Transducers

Current transformers and current transducers—both of which are referred to as CTs—are sensors that measure current. When using CTs, confirm they have the correct output for the meter with which they will be paired.

- Current transformers, which output a current on the secondary wires, can produce dangerously high voltages if the wires are not shunted (that is, shorted, sometimes with a resistor). These CTs are typically rated by the transformer ratio, such as 100:5, where 100 refers to the maximum Amp rating of the primary conducts and 5 refers to the full scale Amp output of the secondary. Connect the leads of this type of CTs to the power meter before placing the CT on load conductors. Wire leads from these CTs must be routed through conduit or contained inside of electrical panels.
- Current transducers, which output a low voltage signal proportional to the current, are intrinsically safe to handle. Short-term power metering equipment typically uses CTs

with a full-scale output of 0.333 Vac. The wires from these CTs do not need to be run in conduit because they are intrinsically safe.

6.1.4.1 Split-Core CTs, Solid-Core CTs, and Current-Only Metering

Solid-core CTs, which have higher accuracy than split-core CTs, are in the shape of a ring, so the wire conductors of the load must be threaded through the center hole. This requires the load to be turned off while the wire is temporarily disconnected.

For temporary metering installations, split-core CTs are recommended to avoid turning off customer loads. Split-core CTs can be opened up and wrapped around a current conductor without shutting down the load. As some accuracy can be lost due to electromagnetic field (emf) leakage at the core junctions, the CT should have an accuracy of ±1.0% or better and a phase angle shift of 2° or less.

When only current metering is required, CTs with Vdc output (typically 2.5 Vdc) are used with a dc voltage logger.

6.2 Light/Motor/Event

There are several types of event (or status) loggers. Some have specific uses, such as light on/off loggers; others, such as state loggers, can be triggered by various inputs. All of these logger types record a date and time stamp when an event occurs.

6.2.1 Light

Light on/off loggers use a photo sensor with a sensitivity adjustment for the threshold setting. This setting triggers an event when the light level transitions above or below the threshold level.

6.2.2 Motor

Motor on/off loggers sense an electromagnetic field to trigger an event when the emf transitions above or below a threshold. The emf that triggers an event can be from a motor, a coil winding on a valve, or a conductor separated from other phase conductors.

6.2.3 State

State loggers record either the state of a switch or the open or closed position of a door or valve. A one-second time resolution on the event is typical for these types of loggers.

6.3 Temperature

Temperature is measured using a thermometer and there are several sensor types in use, such as:

- **Resistive temperature devices (RTD):** Available in various temperature ranges, RTDs are generally used in combination with a meter specifically designed for that type of sensor. Metal RTDs (such as platinum) generally have linear resistance with temperature. Thermistors, which are the most common RTD, have a ceramic semiconductor base and an electrical resistance that drops non-linearly with temperature.
- **Thermocouples:** Two dissimilar metals joined at the tip of a probe produce a very small voltage proportional to the temperature. A junction at a reference temperature is required. Types T, J, and K are common thermocouples suitable for different temperature ranges.

- **Integrated circuit (IC):** A semiconductor chip with a current that is linear with temperature characteristics.
- **Infrared (IR):** As infrared radiation is emitted by all objects, the peak emitted wavelength is correlated with a black body distribution curve to determine the temperature. An IR is a non-contact device.

Temperature sensors are connected to—or contained within—a meter that converts the sensor signal into a temperature reading. The ideal resolution of the temperature meter or logger is determined by the temperature range:

- Temperature measurements ranging from 32°F to 120°F should have a logger or meter with a resolution of 0.1°F and an accuracy of ±1°F or better.
- Temperature measurements ranging from 100°F to 220°F should have a resolution of 0.5°F and accuracy of ±2°F or better.
- Temperature measurements above 220°F should have a resolution of 1°F and an accuracy of ±4°F.

6.3.1 Loggers with Internal Probes

Many small battery-operated temperature loggers are available; however, as the sensor for such loggers is typically located within the case, these loggers are generally only suitable for air temperature measurements.

6.3.2 Loggers with External Probes

Temperature loggers having external probes are required for surface mountings, liquid immersion, or small openings into air streams. For any application in which the sensor may become damp or wet, use an encapsulated probe. Probes in stainless steel sheaths will typically not be compromised by harsh environments.

6.3.3 Differentials

Measurements used to estimate differential temperature—such as supply and return air—should use a matched pairs of sensors.

6.4 Humidity

Humidity can be measured using either a humidity sensor connected to an analog signal logger or a humidity meter. Many humidity meters also meter dry-bulb temperature and can display other humidity-related values. The humidity measurement should have a resolution of 0.1% RH and an accuracy of ±2.5% RH or better over a range from 10% to 90% RH.

As humidity sensors become saturated easily and remain so for a period longer than the air is saturated, avoid condensation conditions.

6.5 Pressure

Pressure measurement instruments are categorized for use with high-pressure liquids/gases or low-pressure gases. Recording these measurements typically requires the use of a pressure sensor wired to an analog input recorder. Pressure sensors typically have 4-20 mA or 0-5 Vdc output.

6.5.1 High-Pressure Sensors

These sensors are used for refrigerant systems, compressed air, water storage, or water pumping. The common unit of measure is pounds-per-square-inch gauge (psig), which is the pressure above ambient atmospheric pressure. These pressure measurements should have a resolution of 1 psig and an accuracy of ±1% or better.

6.5.2 Low-Pressure Sensors

These sensors are used for barometric readings, air ducts, and combustion exhaust pipes, and one type—a differential pressure sensor—is routinely used to measure duct static pressure. The common unit of measure is inches of water column (IWC). The low-pressure measurements should have a resolution of 0.1 IWC and an accuracy of ±1% or better.

6.5.3 Instantaneous

Use digital pressure gauges for conducting instantaneous readings.

6.6 Flow

The majority of flow measurements will be for water (liquid), air (gas), or natural gas. Flow measurement accuracy is particularly dependent on the proper use of the flow instruments.

6.6.1 Water

Water flow instruments should have an accuracy of ±2% or better of full-scale flow rate.

- Paddle wheels and turbines are commonly used water flow sensors, but they must be inserted into the flow.
- Ultrasonic flow meters use pulsed sound signals applied to the outside of the pipe. These signals measure water-based liquids in pipes without interrupting the flow to install the meter.

6.6.2 Air

Measurements may be taken of conditioned air or exhaust, and the measurement should have an accuracy of ±5% or better. Hot-wire anemometers, pitot tubes, calibrated duct fans, balometers, and capture hoods are instruments used for air velocity or volume flow rates.

6.6.3 Natural Gas

Natural gas meters, which use a positive displacement approach to measure flow, should be installed inline. These meters should have an accuracy of ±1% or better and be temperature-compensated.

6.7 Other Sensors

Other commonly used sensors and meters are these:

- Occupancy sensors
- CO_2 sensors
- Combustion gas analyzers

- Solar radiation sensors (such as pyranometers)
- Wind speed sensors
- British thermal unit (Btu) meters.

When selecting the desired level of accuracy for each of the sensor types, consider both cost-effectiveness and the importance of the measurement to the final savings estimations.

6.7.1 Digital Cameras

Digital cameras are very useful in documenting metering equipment before and after its installation and during the evaluation of the EEM.

6.8 Pulse and Analog Signal Loggers

Certain data loggers (single channel and, more commonly, multi-channel) record generic sensor signal inputs. Depending on the logger, digital channels or pulse loggers can be used to count (1) pulses, (2) switch openings and closings, and (3) the percentage of time a switch is open or closed during an interval.

Inputs are categorized as digital or analog. Analog signal input channels include 4-20 mA, various ranges of dc voltage, and resistance in ohms. The logger should have an accuracy of ±0.5% or better.

Sensor accuracy is a separate measure that is dependent on the type of sensor and should be considered in the final measurement.

6.8.1 Battery Operated

Data loggers may be battery operated, powered by a separate power supply, or powered by a line voltage input. When using battery-operated loggers, ensure that the useful life of the battery is sufficient to allow the unit to remain operational until the next site visit.

The time accuracy of data loggers should be one minute per month or better. Logger and sensor calibration should be conducted as often as the manufacturer suggests; however, review all measurements for validity.

7 Data Storage, Retrieval, and Handling

There is a wide range of commercially available data loggers. When selecting a data device for a project site, consider the data storage specifications and retrieval requirements. Also, it is important to handle the data appropriately after retrieval, which includes making backup copies in the event that original files become corrupted.

7.1 Data Storage

Although the memory storage capacity of data loggers varies widely, loggers ideally will have sufficient capacity to store at least one month of data. Memory time capacity depends on the recording interval, the number of channels active, and the number of parameters stored. However, event logger memory can quickly reach capacity if the trigger condition is met frequently. (This occurs, for example, when there is a short delay time for occupancy sensors on lighting controls.) Review the manufacturer's instructions for details as to how long a logger can record data before the memory reaches capacity.

7.2 Retrieval

Evaluation of EEMs generally entails short-term metering. At the end of the metering period, the logger is retrieved and data are collected by direct connection between the logger and a laptop computer. While the metering equipment is still on site, field evaluation staff should review the data to confirm that (1) all necessary information was collected and (2) the data are within valid ranges.

The data retrieval method depends on the logger, and manufacturers typically have customized software to communicate with the logger. Also, some manufacturers have specialized interface cables to connect the logger to a computer. With some loggers, communication and data retrieval can occur by alternative methods such as modems with landlines or cell phones, Ethernet and Internet, and other digital contact via local networks.

7.3 Handling

After retrieving the data, make backup copies immediately. For the data files, use a filename convention that includes the site, EEM, logger number, and date.

Because data logger software generally stores the raw data in a proprietary format, export a copy of the data into a common format, such as comma-separated value (CSV), ASCII, or Excel. Store the exported data on a secure system that is regularly maintained and monitored.

8 Metering Methods by Load Type

This section provides summary tables of metering methods for various load types and the preferred metering approach for each type. To determine the appropriate metering approach, categorize the characteristics of the load type into one of the following load types defined in Table 1. Use these definitions to find the load type that most closely matches the EEM to be evaluated.

Some measures (such as building envelopes) do not directly use energy, but they impact energy use. In those cases, the end use that would be metered is the energy-using equipment impacted by the measure. In general, these categories are listed in increasing order of metering complexity. The example end uses provided in tables 2 through 7 are not intended to be an exhaustive list of measures; rather they are a guide for the most common energy efficiency measures. The examples are predominantly electric loads, because they account for the most commonly evaluated measures.

Table 1. Load Type Definitions

Load Type	Definition
Constant Load Time-Dependent	The load or energy demand does not change. The energy use depends only on when the load is operated, and there is a schedule of operation.
Constant Load Cycling	The load or energy demand does not change. The energy use depends only on when the load is operated, and conditions dictate when the load cycles on or off.
Variable Load Weather-Dependent	The load or energy demand varies with the weather and does not run constantly.
Variable Load Continuous	The load or energy demand varies, and the equipment runs continuously during a scheduled period.
Variable Load Cycling	The load or energy demand varies. The load may (1) be repetitive, (2) turn on and off, or (3) cycle based on conditions.
Loads Measured Indirectly	The load or energy demand of the end use cannot be measured directly, so it is calculated from one or multiple metered measurements.

Alternatively, follow the flowchart in Figure 1.

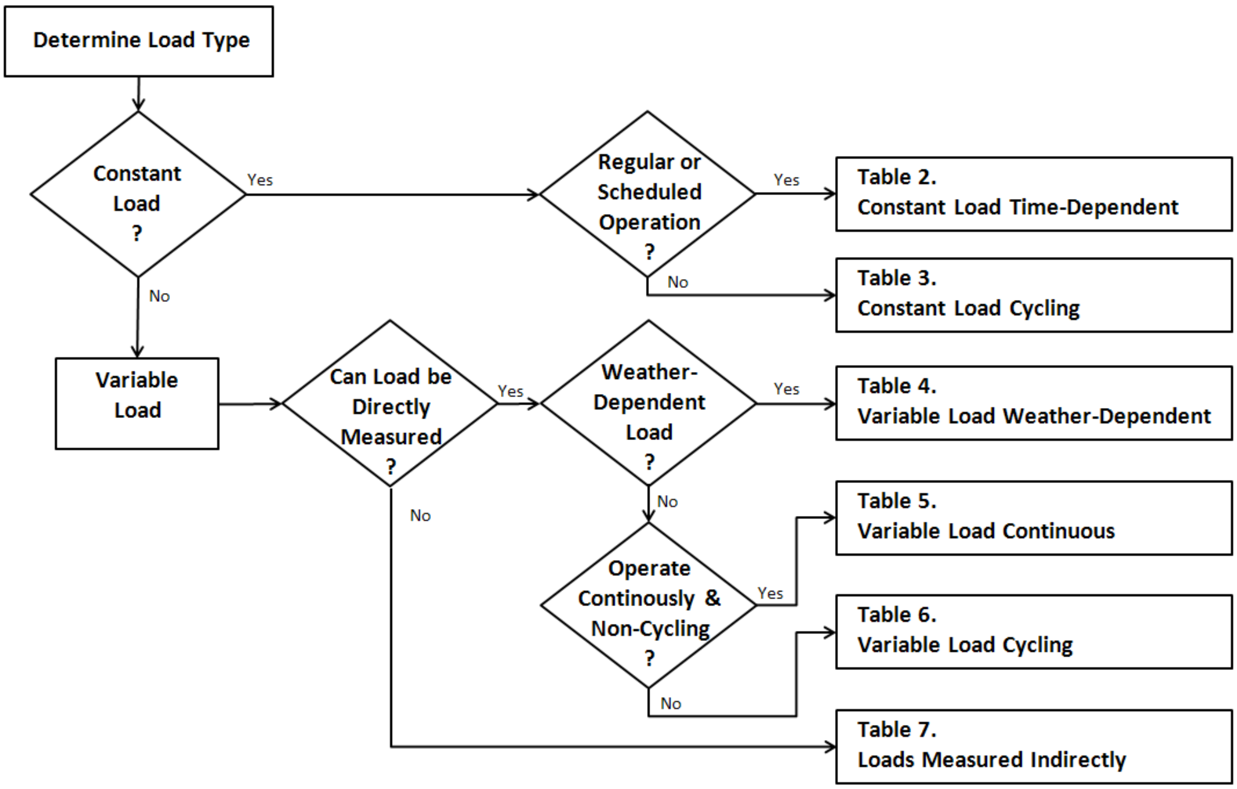


Figure 1. Load type determination flowchart

8.1 Levels of Rigor

Rigor is associated with the level of precision, with a higher level of rigor corresponding to a higher level of precision—and, often, with higher costs or more labor hours. Because the level of rigor varies widely among metering methods, consider this relationship between precision and cost when selecting the preferred metering approach.

Typically, there are multiple metering methods possible for the majority of load types, so the metering methods shown in tables 2 through 7 are ranked by level of rigor. Of the three levels of rigor, Level 1 is the lowest level and Level 3 is the highest level of rigor.

Identify the preferred level of rigor when developing the measurement approach. The tables list alternative levels of rigor that may be selected for the measurement approach if circumstances justify the level selection. The durations listed in tables are minimum monitoring times, but M&V plans may request longer periods or multiple periods with different conditions. Conditions may include various seasons for weather-dependent loads or periods with different operating hours (such as in schools or colleges). Selecting when monitoring occurs can be as important (or more important) than the duration of the monitoring.

Current (or Amp) metering rather than power metering can be conducted when:

- A load has a stable or well-defined power factor and the interval of recording is short relative to the system cycle

- Metering is done only to determine operating hours.

With Amp metering, additional analysis effort is needed to convert current data to power rather than directly metering power.

8.2 Proxy Measures

Indirect measurement of energy is the most practical approach for many end uses, and there are suitable substitute proxy measurements for most end uses. Proxy measurements generally produce less accurate results than direct measurements. Most proxy measurements require a multiplier or scalar factor, which is either measured or determined. As an example, a natural gas-fired boiler with a constant burner flow rate can be measured by metering the "on" status of the combustion air fan, which is energized when the burners are operating. Alternatively, the burner gas flow rate can be measured by using the utility gas meter and a stopwatch, if all other gas appliances are switched off.

Table 2. Constant Load Time-Dependent

Example End-Use	Rigor Level
Lighting (non-dimming) Pool pumps Constant-speed chilled water pumps Condenser water pumps Constant volume fan motors Data center equipment	**Level 1—Preferred Approach** Equipment: On/off loggers Additional measurement: Instantaneous Volts, Amps, kW, and power factor (if wattage is not deemed) Duration: Two weeks Interval: n/a
	Level 2 Equipment: Amp metering Additional measurement: Instantaneous Volts, Amps, kW, and power factor. Duration: Two weeks Interval: 5 minutes
	Level 3 Equipment: Power (kW) metering Duration: Two weeks Interval: 15 minutes

Table 3. Constant Load Cycling*

Example End-Use	Rigor Level
Lighting with occupancy sensors or bi-level controls Refrigerators and freezers Water heaters, electric Plug-in loads Household and office electronics Electronically commutated motor fans Electric ovens or grills	**Level 1** Equipment: On/off loggers Additional measurement: Instantaneous Volts, Amps, kW, and power factor (if wattage not deemed). Duration: Two weeks Interval: n/a
	Level 2—Preferred Approach Equipment: Amp metering Additional measurement: Instantaneous Volts, Amps, kW, and power factor. Duration: Two weeks Interval: Two minutes
	Level 3—Preferred Approach Equipment: Power (kW) metering Duration: Two weeks Interval: 15 minutes

* Either meter for operating hours or have well-defined power factor profiles.

Table 4. Variable Load Weather-Dependent

Example End-Use	Rigor Level
Air conditioner* Heat pump* Packaged HVAC Chiller Cooling tower* Refrigeration Furnace, electric	**Level 1** For those indicated by (*) and applied only for single compressor/motor w/no VSD Equipment: On/off loggers, outdoor temperature logger Additional measurement: Instantaneous Volts, Amps, kW, and power factor (if wattage not deemed) Duration: one month Interval: n/a
	Level 2—Preferred Approach For loads without VSDs Equipment: Amp metering, outdoor temperature logger Additional measurement: Instantaneous Volts, Amps, kW, and power factor Duration: One month Interval: Two minutes
	Level 3—Preferred Approach Equipment: Power (kW) metering, outdoor temperature logger Duration: One month Interval: 15 minutes

Table 5. Variable Load Continuous

Example End-Use	Rigor Level
Water pump with VSD Warehouse lighting with daylight dimming Lighting with dimming controls Air compressor with VSD Fan with VSD Motor with VSD Industrial Process Equipment Boiler*	**Level 1 – N/A**
	Level 2 Equipment: Amp metering, (*gas meter with pulse output and pulse logger) Additional measurement: Instantaneous Volts, Amps, kW, and power factor at five different speeds or conditions. Duration: Four weeks Interval: Two minutes
	Level 3—Preferred Approach Equipment: Power (kW) metering, (*gas meter with pulse output and pulse logger) Duration: Four weeks Interval: 15 minutes

Table 6. Variable Load Cycling

Example End-Use	Rigor Level
Air compressor Injection molding machines* Oil well pumpjack* Industrial Process Equipment	Level 1 – N/A
	Level 2 Equipment: Amp metering Additional measurement: Instantaneous Volts, Amps, kW, and power factor. Duration: Four weeks (*Two weeks) Interval: 2 minutes
	Level 3—Preferred Approach Equipment: Power (kW) metering Duration: Four weeks (*Two weeks) Interval: 15 minutes

Table 7. Loads Measured Indirectly

Example End-Use	Example of Preferred Approach When Direct Measurement Not Practical
Furnace, gas Boiler, gas Water Heater, gas	Duration: One month Interval: 15 minutes **For Constant Rate Burners** Equipment: On/off motor loggers mounted on gas valve 24 Vac coil or combustion air fan motor Additional measurement: Measure burner flow rate using utility meter with all other loads off and stopwatch **For Variable Rate Burners** Equipment: Amp metering of combustion air fan or analog signal logger for modulating valve Additional measurement: Measure burner flow rate using utility meter with all other loads off and stopwatch for three typical flow-rate conditions and correlate to fan Amps or valve signal
High voltage loads >600 Vac (such as 4,160 Vac motors or chillers)	Equipment: Amp metering on 5 Amp secondary of CT used for panel mount display of load Amps Determine CT ratio: kW=V*A* (Assume V and) Duration: One month Interval: 15 minutes

9 Resources

Alereza, T.; Martinez, M.; Mort, D. (1989). "Monitoring of Electrical End-Use Loads in Commercial Buildings." *ASHRAE Transactions*. (95:2). http://repository.tamu.edu/bitstream/handle/1969.1/6558/ESL-HH-88-09-52.pdf?sequence=3.

American Society of Heating Refrigeration and Air-Conditioning Engineers (ASHRAE). (1995). *ASHRAE Handbook: HVAC Applications.* Chapter 37: "Building Energy Monitoring."

ASHRAE. (1996). *Methodology Development to Measure In-Situ Chiller, Fan, and Pump Performance*. ASHRAE Research Report, Research Project 827. http://industrycodes.com/products/4a6765/rp-827-methodology-development-measure-situ.

ASHRAE. (June 2002). *Guideline 14-2002*: *Measurement of Energy and Demand Savings*. http://gaia.lbl.gov/people/ryin/public/Ashrae_guideline14-2002_Measurement%20of%20Energy%20and%20Demand%20Saving%20.pdf.

ASHRAE. (2010). *Performance Measurement Protocols for Commercial Buildings*.

Bonneville Power Administration. (September 2010). *End-Use Metering Absent Baseline Measurement: An M&V Protocol Application Guide, Version 1.0*. www.bpa.gov/energy/n/pdf/BPA_End_Use_Metering_Absent_Baseline_Measurement_A_Measurement_and_Verification_Protocol_Application_Guide.pdf.

California Public Utilities Commission (CPUC). (June 2004). *The California Evaluation Framework*. www.calmac.org/publications/California_Evaluation_Framework_June_2004.pdf.

Dent, C.; Mort, D., "Electrical Metering Fundamentals", (June 1-3, 1992). "Metering Your End-Use Needs". Bend, Oregon.

Federal Energy Management Program (FEMP). (September 2000). *Measurement & Verification Guidelines for Federal Energy Projects Version 2.2*. DOE/GO-102000-0960. www.nrel.gov/docs/fy00osti/26265.pdf.

FEMP. (April 2008). *Measurement & Verification Guidelines for Federal Energy Projects Version 3*. www1.eere.energy.gov/femp/pdfs/mv_guidelines.pdf.

FEMP. (August 2011). *Best Metering Practices Release 2.0*. www1.eere.energy.gov/femp/pdfs/mbpg.pdf.

Institute of Electrical and Electronics Engineers (IEEE). (1989). *Master Test Guide for Electrical Measurements in Power Circuits*. ANSI/IEEE Std. 120-1989.

International Performance Measurement and Verification Protocol (IPMVP). (2010). IPMVP *Volume 1: Concepts and Options for Determining Energy and Water Savings*. EVO 10000 – 1:2010. Washington, D.C.: Efficiency Valuation Organization. www.evo-world.org/index.php?option=com_form&form_id=38.

National Electric Code (NEC). (2011). National Fire Protection Association (NFPA) 70.

National Institute for Occupational Safety and Health (NIOSH). (July 1999). *Preventing Worker Deaths from Uncontrolled Release of Electrical, Mechanical, and Other Types of Hazardous Energy*. Publication No. 99–110. www.cdc.gov/niosh/docs/99-110/pdfs/99-110.pdf.

Occupational Safety and Health Administration (OSHA). *Standard 1910* (emphasis on Subpart S – Electrical). www.osha.gov/pls/oshaweb/owasrch.search_form?p_doc_type=STANDARDS&p_toc_level=1&p_keyvalue=1910.

Chapter 10: Peak Demand and Time-Differentiated Energy Savings Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Frank Stern and Justin Spencer
Navigant
Boulder, Colorado

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68566
October 2017

Contract No. DE-AC36-08GO28308

Chapter 10: Peak Demand and Time-Differentiated Energy Savings Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in April 2013. The content in this version has been updated.

Frank Stern and Justin Spencer
Navigant
Boulder, Colorado

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68566
October 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgements

The chapter authors wish to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Richard Spellman, GDS
- Sami Khawaja, Cadmus
- Steve Patterson, CDP Strategies
- Pete Jacobs, Building Metrics.

Suggested Citation

Stern, F.; Spencer, J. (2017). *Chapter 10: Peak Demand and Time-Differentiated Energy Savings Cross-Cutting Protocol. The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68566. http://www.nrel.gov/docs/fy17osti/68566.pdf

Acronyms

AMI	automated metering infrastructure
CFL	compact fluorescent lamp
DOE	U.S. Department of Energy
EPA	U.S. Environmental Protection Agency
HVAC	heating, ventilating, and air conditioning
IPMVP	International Performance Measurement and Verification Protocol
ISO	independent system operator
kWh	kilowatt hour
LOLP	loss of load probability
MW	megawatt
MWh	megawatt hour
NEEP	Northeast Energy Efficiency Partnerships
NILM	non-intrusive load monitoring
NREL	National Renewable Energy Laboratory
RLF	rated load factor
RTO	regional transmission organization
SAE	statistically adjusted engineering
TRM	technical reference manual
UMP	Uniform Methods Project

Protocol Updates

The original version of this protocol was published in April 2013. This version has been updated to incorporate the following revisions:

- Expanded definition of coincidence and diversity factors
- Expanded interval metered data analysis discussion to include recent work related to non-intrusive load monitoring
- Added discussion of preferred applications and limitations
- Added recent relevant references.

Table of Contents

List of Figures

List of Tables

1 Introduction

Savings from electric energy efficiency measures and programs are often expressed in terms of annual energy and presented as kilowatt-hours per year (kWh/year). However, for a full assessment of the value of these savings, it is usually necessary to consider the measure or program's impact on peak demand as well as time-differentiated energy savings.[1]

This cross-cutting protocol describes methods for estimating the peak demand and time-differentiated energy impacts of measures implemented through energy efficiency programs.[2]

[1] While natural gas peak demand impacts can be important in some situations, most utility programs do not attempt to generate peak demand impacts for natural gas or any other energy type except for electricity. As a result, this protocol focuses only on electricity peak demand and time-differentiated energy savings. The fundamental principles laid out in this chapter may be applied to peak demand and time-differentiated savings for other energy and water impacts.

[2] As discussed in the "Considering Resource Constraints" section of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

2 Purpose of Peak Demand and Time-Differentiated Energy Savings

Energy efficiency measures and programs frequently reduce peak demand and, consequently, the need for investment in new generation, transmission, and distribution systems. To estimate the value of these avoided costs—called "avoided capacity costs"—it is necessary to estimate peak demand savings. Peak demand savings are typically expressed as the average energy savings during a system's peak period. Avoided capacity costs can account for a substantial portion of the total value of an energy efficiency measure or program, particularly for those that produce savings coincident with the system peak.

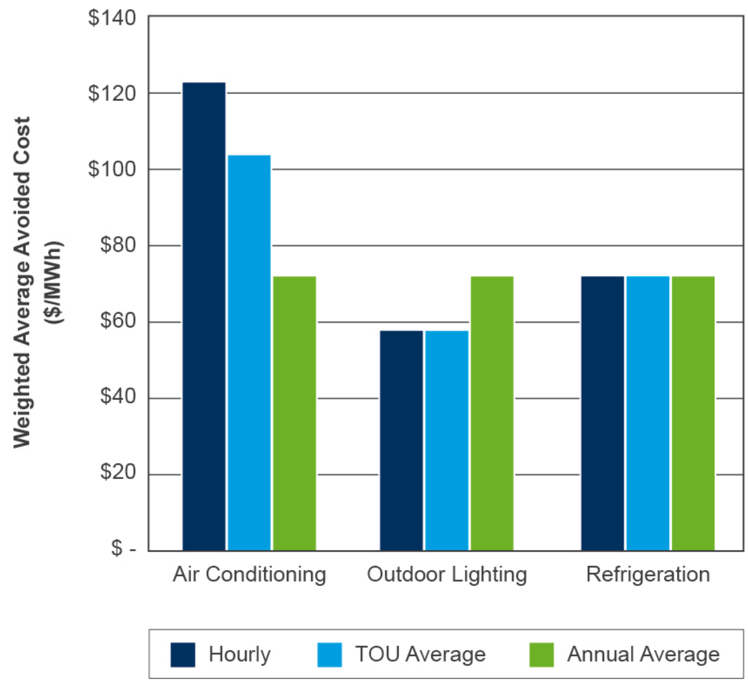
The need to estimate peak demand savings is becoming more important as the value of avoided capacity costs increases, and as regional transmission organizations (RTOs, such as PJM and independent system operator [ISO]-New England) allow energy efficiency resources to be bid into the forward-capacity markets and be used to earn revenues.[3]

In addition to considering peak demand savings, evaluators often must calculate time-differentiated energy savings, or the energy savings that occur at different times of the day (e.g., morning or evening) or times of the year (e.g., summer or winter). This is because avoided energy costs are typically provided in terms of costing periods, which allocate the 8,760 hours of the year into periods with similar avoided capacity costs. These costing periods, which are utility/RTO/ISO-specific, tend to vary monthly, seasonally, and/or in terms of time of day (peak, off-peak, super-peak).[4]

When estimating the impacts of energy efficiency measures and programs, calculating load impacts on an hourly basis provides flexibility in applying the results to a variety of costing period definitions. The cost period used can significantly affect the value of the energy savings. For example, a measure that reduces energy mostly at night (typically a low-cost, off-peak costing period) is not as valuable as one that reduces energy mostly during the summer afternoon peak load periods, as shown in Figure 1.

[3] Regional transmission markets obtain the capacity resources for ensuring system reliability. In some regions, energy efficiency is considered to be a resource comparable to traditional generating resources and can be included into these markets on an equivalent basis to supply-side resources. Bids must be supported by measurement and verification.

[4] Avoided energy costs tend to be more expensive during periods of higher demand because generating units available during those times tends to have lower efficiency and higher operating costs.



Source:

Figure 1. Consideration of time-differentiation in energy savings significantly affects estimates of the value of savings (EPA and DOE 2006)

In this example, the air conditioning efficiency measure has a higher value when considering hourly savings and costs because usage is higher when avoided costs are more expensive. Conversely, an outdoor lighting measure will have a lower value when considering hourly savings and costs as those savings typically occur during off-peak nighttime hours.

Peak periods typically relate to capacity limitations on physical equipment within the grid and may vary by location (e.g., a specific overloaded feeder may have a different peak than the overall grid). The load will approach the available capacity as a result of a combination of weather and behavior, causing the load to increase or some other impact causing the available capacity to decrease (such as a shortage of natural gas). A peak period can range from one 5-minute burst at a random time, driven by an irregular high-usage event,[5] to one hour per year or several hours per day during a season. When transmission and distribution operators define peaks, they frequently use a combination of time of day and weather as peaks tend to occur during extreme hot or cold temperatures.

[5] In the UK, short bursts of electricity peak demand have correlated with the end of events, such as popular TV shows, World Cup matches, or solar eclipses.

3 Key Concepts

Understanding demand savings requires understanding the relationship between several factors—some of which have conflicting definitions. Figure 2 shows one such construct.

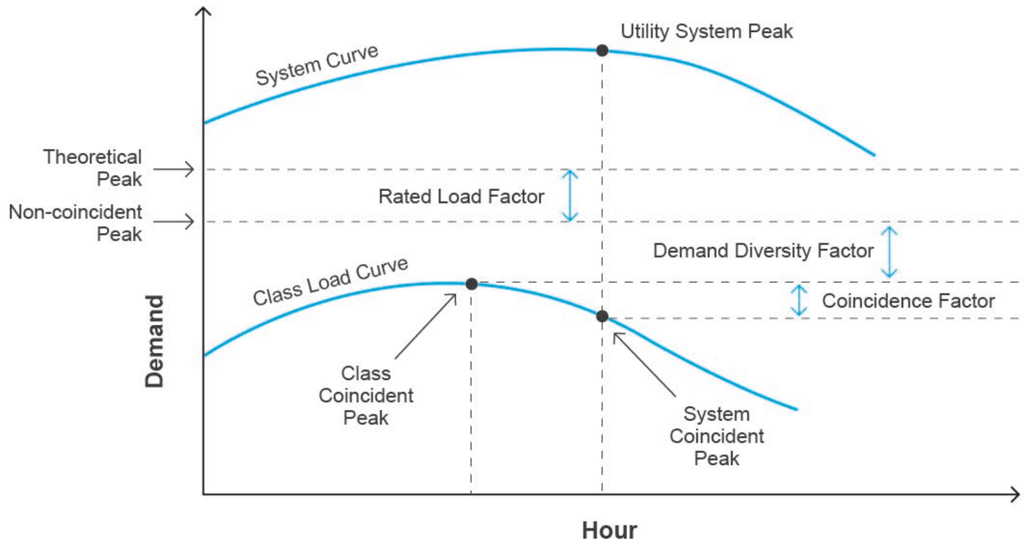


Figure 2. Demand savings relationships[6] (Jacobs 1993)

These brief definitions describe the key factors within this construct:

- **Peak period** is the period during which peak demand savings are estimated. Some utilities have a winter and summer peak period. The peak period definition may also include weather conditions.
- **Theoretical peak** is the usage of a population of equipment if all operate at nameplate capacity.
- **Non-coincident peak** is the sum of the individual maximum demands, regardless of time of occurrence within a specified period.
- **Rated load factor (RLF)** is the ratio of maximum operating demand of a population of equipment to the nameplate power/capacity. It is the ratio of non-coincident peak to theoretical peak. For example, a building that dims its lamps to 90% of their output has a RLF of 0.9.
- **Demand diversity factor** is the ratio of the peak demand of a population of units to the sum of the non-coincident peak demands of all individual units. While an individual efficiency technology may save a certain amount of demand, those technologies do not all operate at the same time across all buildings throughout the

[6] Rated load factor, demand diversity factor, and coincidence factor are sometimes combined and referred to as "coincidence factor."

region. For example, if a maximum of 7 of 10 installed compact fluorescent lamps (CFLs) are on at any given time, the diversity factor is 0.7.

- **Coincidence factor** is the fraction of peak demand of a population in operation at the time of system peak. Thus, it equals the ratio of the population's demand at the time of the system peak to its non-coincident peak demand. The peak demand use for a given building and end use typically do not align exactly with the utility system peak, which is how avoided peak demand is defined. For example, if at the time of system peak, only 3 of the 7 CFLs mentioned above are on, then the coincidence factor is 3/7.

The Northeast Energy Efficiency Partnerships (NEEP) defines a coincidence factor as, "The ratio of the average hourly demand during a specified period of a group of electrical appliances, or consumers to the sum of their individual maximum demands (or connected loads) within the same period." (NEEP 2011). This corresponds to the product of rated load factors, demand diversity factors, and coincidence factors, as defined above.

IEC60050—International Electrotechnical Vocabulary (International Electrotechnical Commission 2016) defines coincidence and diversity as:

- **Coincidence factor** is the ratio, expressed as a numerical value or as a percentage, of the simultaneous maximum demand of a group of electrical appliances or consumers within a specified period to the sum of their individual maximum demands within the same period. Per this definition, the value always remains less than or equal to 1 and can be expressed as a percentage.

- **Diversity factor** is the reciprocal of the coincidence factor, which means it will always be greater than or equal to 1.

The following terms are also important in understanding peak demand:

- **Average (or Annual Average) megawatt (MWa or aMW).** One megawatt of capacity produced continuously over a period of one year.

  1 aMW = 1 MW x 8,760 hours/year = 8,760 MWh

- **Load factor.** The ratio of average energy savings to peak energy savings. This is also known as "peak coincidence factor" (NYSERDA 2008).

$$Load\ factor = \frac{Energy\ savings}{Peak\ demand\ savings \times 8760\ hours}$$

- **Loss of load probability (LOLP).** The likelihood that a system will be unable to meet demand requirements during a certain period. LOLP can be used to distribute avoided capacity costs to each hour of the year.

4 Methods of Determining Peak Demand and Time-Differentiated Energy Impacts

Estimating peak demand and time-differentiated energy savings may require different techniques than estimating annual energy savings. For example, the method used to estimate demand savings may not be the most appropriate method to estimate energy savings—and vice versa (Fels 1993).

Peak demand and time-differentiated energy impacts are typically more difficult to measure than annual energy savings impacts (York 2007), and may require additional metering or simulation analysis to estimate these impacts accurately.

Peak demand savings and time-differentiated energy savings can be estimated with:

- Engineering algorithms
- Calibrated hourly building simulation modeling
- Billing data analysis
- Interval metered data analysis
- Non-intrusive load monitoring
- End-use metered data analysis
- Survey data on hours of use.

Approaches can also be combined to leverage available information. For example, a method used to estimate annual energy savings that does not directly provide peak demand savings (such as monthly billing data analysis) can be used with load shapes to estimate peak or hourly impacts.

The more closely the time-based impacts of a device or measure can be calculated, the more accurate and supportable the results will likely be. This means that directly measuring an end use of interest will generally provide the most accurate results. Other methods are generally used to provide a lower-cost means of estimating measure-level peak demand impacts or, in rare cases, are used where direct measurement does not work.

The following sections examine the various methods for estimating peak demand and time-differentiated energy savings, and discuss the preferred applications and limitations for each method. Table 1 presents a summary of all approaches.

4.1 Engineering Algorithms

Algorithms can be used to estimate peak demand savings. The demand algorithm in Equation 1 is similar to the energy algorithm in Equation 2 (used to estimate annual energy savings), except that the demand equation has the diversity factor and the coincidence factor in place of the full load hours.

Equation 1

$$\Delta kW_{gross} = units \times RLF \times \left[\left(\frac{kW}{unit}\right)_{base} - \left(\frac{kW}{unit}\right)_{ee} \right] \times \boldsymbol{DF} \times \boldsymbol{CF} \times (1 + HVAC_d)$$

Where:

ΔkW_{gross}	=	gross demand savings
Units	=	units of measure installed in the program
RLF	=	rated load factor
kW/unit	=	unit demand of measure
DF	=	diversity factor
CF	=	coincidence factor
$HVAC_d$	=	HVAC system interaction factor for demand

Source: (TecMarket Works 2004)

Equation 2

$$\Delta kWh_{gross} = units \times RLF \times \left[\left(\frac{kW}{unit}\right)_{base} - \left(\frac{kW}{unit}\right)_{ee} \right] \times \boldsymbol{FLH} \times (1 + HVAC_c)$$

Source: (TecMarket Works 2004)

Preferred Applications: Engineering algorithms are most useful for quickly and inexpensively estimating prescriptive program impacts for measures or programs with accurate and stable parameter estimates from previous studies. This approach is common for residential and commercial lighting measures where confidence and diversity factors have been previously measured.

Limitations: To acquire good estimates for an alogrithm's time-based parameters, results from another method must be extrapolated from another place or a previous study. Differences in the installation context may cause biased results. For example, the residential lighting consumption patterns of retirees in Arizona may be very different than the residential lighting consumption patterns of young apartment-dwellers in Brooklyn.

4.2 Calibrated Hourly Building Simulation Modeling

Hourly building simulation modeling (International Performance Measurement and Verification Protocol [IPMVP] Option D) can produce hourly savings estimates for whole buildings as well as for specific end uses. It is an excellent means of estimating peak demand and time-differentiated energy savings. A building energy simulation model combines building characteristic data and weather data to calculate energy flows. While hourly models calculate energy consumption at a high frequency, non-hourly models may use simplified monthly- or annual-degree-day or degree-hour methods with limited applications to calculating time-differentiated impacts. Hundreds of building energy simulation programs have been developed over the past 50 years (Crawley 2005). This chapter differentiates between calibrated building simulations, which calibrate models to the primary usage data from the group studied, and uncalibrated building simulation, as described in Section 5.3. Calibrated building simulation is generally much more accurate than uncalibrated building simulation.

Preferred Applications: Building simulation models serve as an ideal method for extrapolating observable data to unobservable scenarios, or for capturing interactive effects that may be difficult or impossible to measure directly. They prove most applicable for heating, ventilating, and air conditioning (HVAC) and shell measure impacts. Preferred applications include estimating savings for new construction applications that have unobservable code-baseline buildings, extrapolating measured weather-sensitive data to extreme weather conditions, and calculating HVAC's interactive effects on lighting and equipment measures.

Limitations: Simulation modeling requires an experienced modeler who understands energy engineering. Using this method does not necessarily provide an estimate of diversified demand. If using a single, typical building, demand savings would be overstated due to a lack of consideration of diversity, which tends to smooth out the usage spikes seen in individual buildings. Consideration of diversity requires either using average schedules in an aggregate building simulation model[7] or simulating a sample of buildings with different sizes, climates, and schedules. A building simulation model, however, can be only as good as the data it calibrates to. Deriving accurate hourly load shape results from a building simulation model requires—at a minimum—calibrating the model to hourly whole-facility load shape data.

4.3 Billing Data Analysis with Load Shapes

Billing data analysis (IPMVP Option C) is a common evaluation method for many common energy efficiency measures and programs and has been used traditionally to develop monthly and annual estimates of energy savings (Agnew 2017). Such analysis entails statistical comparison of pre- and post-participation and/or participant and nonparticipant billing data and may require control for non-programmatic influences, such as weather and economic conditions, to estimate savings.

Billing data analysis does not directly estimate peak demand or hourly energy savings, but can be combined with other tools to estimate these impacts.

Preferred Applications: Billing data analysis can be used to derive a realization rate using an engineering algorithm for energy savings from *in situ* baseline measures, which also may be applied to a demand savings algorithm by using a previously calculated load shape. In this case, the billing analysis would not determine the shape of savings, which derive from the previously defined shape or demand parameters, but it would prove useful for determining the savings' amplitude.

Limitations: As addressed in UMP *Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol*, billing analysis does not work with non-retrofit baseline measures. Further, monthly building peak demand typically does not prove useful for estimating peak demand savings, given that billing peak demand typically is not coincident with utility system

[7] An aggregate building simulation model uses all empirical relationships included in a building simulation model; however, it is calibrated to match the load shape and characteristics of a large group of buildings (e.g., program participants in a given sector) rather than an individual building. Aggregate building simulation models can look more abstract than engineers prefer, while lacking the mathematical purity desired by econometricians accustomed to abstraction.

peak. Isolating the impacts of a specific measure can be difficult because the meter measures usage for an entire building.

4.4 Interval Metered Data Analysis

Utility revenue interval meters, including automated metering infrastructure (AMI) meters, can measure usage at the whole-building level in increments of 15 minutes or less. Because consumption during different periods may be billed at different rates, these meters provide a means for analyzing a customer's load pattern. Interval meter data analysis is essentially the billing data analysis discussed above but with a finer time resolution. As with billing analysis, isolating a specific measure’s impacts can be difficult, and statistical analysis may be required to control for non-programmatic influences. With the advent of AMI and increased access to hourly information, additional statistical approaches (such as conditional demand type analysis on hourly data) can be used to help develop estimates of demand savings.

Preferred Applications: Use of interval metered data for calculating demand savings is preferred for measures with energy savings calculated using billing analysis. This includes retrofit measure baselines with sufficient savings and participants to produce statistically significant results. Large home performance programs with insulation and air sealing measures are a good example application.

Limitations: Isolating a specific measure’s impacts can be difficult, and statistical analysis may be required to control for non-programmatic influences.

4.5 Non-Intrusive Load Monitoring (NILM)

Numerous hardware and software vendors have made varying claims regarding the ability to disaggregate whole-premise interval data into individual end use load shapes—a process known as non-intrusive load monitoring (NILM). NILM combines software and hardware in many different configurations, such as (Mass Save 2016):

1. Software-based solutions that disaggregate data from third-party software
2. Utility smart meters providing data transmission via a Wi-Fi-connected gateway
3. Current transformer-based devices that monitor voltage and current at the home electrical panel or meter
4. Utility meter-reading devices—devices installed at the meter and using meter/optical sensors.

The applications range from devices that record and process data at rates higher than 1 hertz to software working with large volumes of hourly AMI data. The former might cost around $1,000 or more per site, while the latter might cost less than $1 per site.

Preferred Applications: NILM works better for disaggregating larger and more unique loads within smaller facilities,[8] and for loads correlating with other observable data (e.g., weather-sensitive loads such as central air conditioners). Given further NILM advances and declining costs for interval metered data, evaluators should consider using NILM in combination with end-use metering data, if requiring high accuracy. Future NILM improvements may eventually supplant the need for end-use metered data. Example applications can be seen in Decker (2017) and Elszasz (2017).

Limitations: Given the rigorous testing results thus far, NILM should be considered significantly less accurate than direct end-use metering, presenting some of the same accuracy issues encountered through survey data, engineering assumptions, or uncalibrated building simulations (Baker et al. 2016). NILM methods based on data collected at a higher frequency and/or with real and reactive power tend to have a greater capability for disaggregation. Ultimately, to be considered accurate, the results of NILM analysis for most measures still require calibration to end-use metered data for a reasonable proxy set of facilities. Due to the high degree of diversity among facility equipment and load sizes, NILM calibrated for one set of facilities may not provide high-accuracy results with another set of facilities in a different region.

4.6 End-Use Metered Data Analysis

End-use metering data analysis (IPMVP Option A and Option B) can provide a highly accurate means of estimating peak demand or time-differentiated energy savings. As with billing and interval data analysis, end-use metering data analysis entails a statistical comparison of pre- and post-participation and/or participant and nonparticipant billing data. However, end-use metering eliminates most—if not all—of the difficulty in isolating the impacts of specific measures. As a result, end-use metering is considered the gold standard for providing measure-level peak demand or time-differentiated energy savings estimates. ASHRAE has developed a methodology to derive diversity factors and to provide typical load shapes of office buildings' lighting and receptacle loads using end-use metered data (Abushakra 2001). An example of end-use metered data analysis can be seen in the recent Northeast residential lighting and Maryland commercial lighting studies (NMR 2014, Powanda et al. 2015).

Preferred Applications: For most high-impact measures where the baseline need not be observed (e.g., prescriptive measures), end-use metering provides the preferred approach to developing hourly impacts.

Limitations: Evaluators should consider the following when using end-use meter data:

- End-use metering cannot be used easily for measuring small load changes. While the likelihood of picking up a useful signal at the end-use level is higher than at the whole-premise level, fundamental limitations remain on the size of the signal that can be detected when measuring a change in usage. Interactive effects of lighting measures on HVAC systems generally fit in this category.
- Savings should be normalized for weather and other confounding factors.

[8] NILM efforts have focused on residential applications and small commercial applications to a lesser extent. As the facility (and number of different loads) increases, the complexity increases accordingly.

- Pre-installation meter data are difficult to obtain because of the logistics entailed in coordinating with customers, though they may be recreated in some cases (e.g., turning off variable-speed drives). Without pre-installation data, baseline conditions must be estimated with engineering algorithms.
- Although costs have gone down, end-use metering is costly and should be conducted strategically.
- An impact load shape may be different than a post-participation load shape. For example, lighting control impact shapes differ from the shape of controlled lighting. (End uses have shapes with and without the efficiency measures in place, and the difference is the impact shape.) Determining energy efficiency shapes may require either pre-installation metering or reconstruction of the baseline shape.
- Sampling must be done carefully—see UMP *Chapter 11: Sample Design Cross-Cutting Protocol* (Khawaja et al. 2017). Sampling requirements for peak demand savings can significantly differ from requirements for energy savings due to higher coefficients of variation related to peak demand.
- The evaluator must consider the period over which to collect end-use meter data, including the time of year and duration of metering.

4.7 Survey Data on Hours of Use

Evaluators may conduct hours-of-use surveys to identify the times of day when equipment is used. For example, a survey might ask if residential CFLs are used during the summer from 3 p.m.–6 p.m., a typical period for system peak. If the results indicate that 5% of lights were in use at that time, then the combination of the coincidence and diversity factors would be 5%.

Preferred Applications: In populations with highly diverse usage patterns, surveys offer a cost-effective means of determining the range of usage patterns present in a population. Survey results may be combined with end-use metering in nested sampling designs. Surveys work better with more-informed respondents and with targeted questions. For example, asking facility managers what time the building ventilation systems are on should provide more accurate results than asking homeowners to describe their hot water usage patterns.

Limitations: The survey respondent must be knowledgeable about the operation of equipment in question. Relying on customer perceptions may result in significant inaccuracy and bias. Survey sampling should be done in conjunction with the techniques described in UMP *Chapter 11: Sample Design Cross-Cutting Protocol* (Khawaja et al. 2017).

4.8 Combined Approaches

Applying a combination of approaches facilitates using data from several sources to estimate peak demand or hourly energy savings. For example, billing data may be the best approach for estimating energy savings for a low-income program. Engineering algorithms can be used to develop energy and demand savings for each program participant, and these participant energy savings can serve as the independent variables in a statistically adjusted engineering (SAE) billing analysis (see Chapter 8: *Whole-Building Retrofit with Consumption Data Analysis*

Evaluation Protocol). The realization rate from the SAE analysis can be then applied to the population demand estimate from the engineering model.

Combined approaches also include nested samples, where a smaller number of metered sites are used to calibrate surveys from a much larger population. For example, a sample of 30 metered sites may yield a combined coincidence and diversity factor of 6.1%, while the surveys produce an estimate of 5.0% for the metered sample and 5.5% for the entire survey sample. The ratio of 6.1% to 5.0% would be applied to the 5.5% survey sample estimate, resulting in an adjusted factor of 6.7%.

4.9 Summary of Approaches

Table 1 summarizes the approaches in terms of relative cost and relative potential accuracy. In all cases, the accuracy achieved depends on the quality of the analysis.

Table 1. Summary of Approaches

Approach	Relative Cost	Relative Potential Accuracy	Comments
Engineering Algorithms	Low	Low-Moderate	Accuracy depends on the quality of the input assumptions as well as the algorithm. Appropriate for prescriptive measures with good existing data.
Calibrated Hourly Simulation Modeling	Moderate	Moderate	Input assumptions are again important—garbage in, garbage out; appropriate for HVAC and shell measures and HVAC interaction
Billing Data Analysis with Load Shapes	Moderate	Moderate	Typically not directly useful for peak demand or on/off peak energy analysis, but it can be used to leverage other approaches
Interval Metered Data Analysis	Moderate	High	Interval metered data is not available for many customers; it is becoming more feasible with proliferation of advanced metering infrastructure (AMI). Appropriate for residential retrofit programs with HVAC and shell measures.
NILM	Moderate	Moderate	Considered significantly less accurate than direct end-use metering, but less expensive. Most applicable for residential cooling.
End-Use Metered Data Analysis	High	High	Requires careful sampling and consideration of period to be metered. Most applicable to high impact prescriptive measures.
Survey Data on Hours of Use	Low-Moderate	Low	Only applicable in the rare cases when customers can provide better estimates than other available data.

5 Secondary Sources

Evaluators may choose to rely on the following secondary sources rather than on the primary sources listed above because of budget or time constraints.

5.1 Technical Reference Manuals

Technical reference manuals (TRMs) specify savings or protocols for estimating savings for common energy efficiency measures. Typically, TRMs provide approved estimates of energy and demand savings. These deemed savings are based on a regional average for the population of participants rather than for a particular installation.

Although TRMs often provide industry-accepted values or algorithms for calculating savings, users should not assume that an algorithm is correct because it has been used elsewhere. Mistakes are common and should be expected. Values based on "engineering judgment" should be used cautiously.

5.2 Application of Standard Load Shapes

Load shapes provide information about the distribution of energy consumption or savings over time. For example, a savings or impact load shape will indicate the fraction of savings achieved during a specific time period. An hourly annual electricity load shape using fractions of the total annual energy consumption can be multiplied by the annual consumption to derive the consumption in each hour of the year.

Load shapes can be applied to allocate energy consumption into costing periods. Similarly, load shapes can be applied to estimate peak demand and time-differentiated energy savings from energy impacts.

Load shapes may be derived from metering or simulation. A key resource of load shape data is the California Database for Energy Efficiency Resources (CPUC 2011). NEEP has also catalogued load shape data and conducted primary research on several common energy efficiency measures to develop load shapes specific to the northeast region (NEEP 2016). As with any secondary data, the evaluator must consider the applicability of the shapes when climate-sensitive end uses are involved.

5.3 Uncalibrated Simulation Using Standard Building Reference Models

The U.S. Department of Energy (DOE) publishes standard commercial building reference models (Deru 2011) for 16 commercial buildings in 16 climates (for 256 prototypes in all). In addition, DOE publishes a Building America Benchmark definition (Hendron and Engebrecht 2011) that can be used to generate residential building prototypes. These prototypes may be run with an appropriate weather file and be used as a first guess at hourly consumption shapes in the absence of better data, though the results will likely have significant errors. As noted in Section 4.2, calibrating these models to whole-building interval data for building sectors provides much more accurate results.

6 Future Improvements

Best practices for developing time-based estimates of savings are advancing rapidly. Given further advances in NILM and the declining cost of interval metered data, evaluators should consider using NILM in combination with end-use metered data if high accuracy is needed. Future improvements in NILM and accessibility of interval metered data may eventually supplant the need for as much end-use metered data.

7 References

Abushakra, B.; Sreshthaputra, A.; Haberl, J.; Clairidge, C. (2001). *Compilation of Diversity Factors and Schedules for Energy and Cooling Load Calculations.* ASHRAE Research Project 1093. http://repository.tamu.edu/bitstream/handle/1969.1/2013/ESL-TR-01-04-01.pdf?sequence=1.

Agnew, K.; Goldberg, M. (2017). *Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol. The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68564. http://www.nrel.gov/docs/fy17osti/68564.pdf

Baker, M.; Fuller, M.; Hicks, G.; Rodriguez-Anderson, S.M. (2016). *Can We Find the End Use in Smart Metering Data?* 2016 ACEEE Summer Study on Energy Efficiency in Buildings. http://aceee.org/files/proceedings/2016/data/papers/12_775.pdf

California Public Utility Commission (CPUC). (2011). "Database for Energy Efficient Resources." www.deeresources.com.

Crawley, D.; Hand, J.; Kummert, M.; Griffith, B. (2005). *Contrasting the Capabilities of Building Energy Performance Simulation Programs.* U.S. Department of Energy. http://gundog.lbl.gov/dirpubs/2005/05_compare.pdf.

Decker, T. 2017 (forthcoming). "Duckhunt! Benefits and Risks of Load Disaggregation and End-Use Metering for Determining End-Use Loadshapes." International Energy Program Evaluation Conference.

Deru, M.; Field, K.; Studer, D.; Benne, K.; Griffith, B.; Torcellini, P.; Liu, B.; Halverson, M.; Winiarski, D.; Rosenberg, M.; Yazdanian, M.; Huang, J.; Crawley, D. (2011). *U. S. Department of Energy Commercial Reference Building Models of the National Building Stock.* U.S. Department of Energy. http://www.nrel.gov/docs/fy11osti/46861.pdf

Elszasz, J. 2017. "A Snapshot of NILM: Techniques and Tests of Non-Intrusive Load Monitoring for Load Shape Development." International Energy Program Evaluation Conference.

Fels, M.; Keating, K. (1993). "Measurement of Energy Savings from Demand Side Management Programs in U.S. Electric Utilities." *Annual Review of Energy and the Environment*. (18); pp.57-88.

Hendron, R.; Engebrecht, C. (2011). *Building America House Simulation Protocols*. National Renewable Energy Laboratory. http://www.nrel.gov/docs/fy11osti/49246.pdf

International Electrotechnical Commission. (2016). *Electropedia: The World's Online Electrotechnical Vocabulary.* http://www.electropedia.org/

Jacobs, P.; Arney, M.; Keneipp, M.; Eldridge, M. (1993). *Engineering Methods for Estimating the Impacts of Demand-Side Management Programs.* Pleasant Hill, CA: Electric Power Research Institute.

Khawaja, M.S.; Rushton, J.; Keeling, J. (2017). *Chapter 11: Sample Design Cross-Cutting Protocol. The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68567. http://www.nrel.gov/docs/fy17osti/68567.pdf

Mass Save. (May 2016). "Non-Intrusive Load Monitoring (NILM): Presentation to the EEAC." http://ma-eeac.org/wordpress/wp-content/uploads/Non-Intrusive-Load-Monitoring-Presentation.pdf

New York State Energy Research and Development Authority (NYSERDA). (2008). "Energy Efficiency Portfolio Standard Administrator Proposal."

NMR Group. 2014. "Northeast Residential Lighting Hours-of-Use Study."

Northeast Energy Efficiency Partnerships (NEEP). (June 2016). *Load Shape Catalog Project Summary and Recommendation for Future Research* http://www.neep.org/sites/default/files/LoadshapeCatalog_Future%20Researchv7_final.pdf.

Northeast Energy Efficiency Partnerships (NEEP). (March 2011). *Glossary of Terms: A Project of the Regional Evaluation, Measurement and Verification Forum*. http://neep.org/uploads/EMV%20Forum/EMV%20Products/EMV%20Glossary%20of%20Terms%20and%20Acronyms%20-%20Version%202%20FINAL.pdf.

Powanda, P., Lisle, H., Spencer, J., Rivas, J. "Getting Over the Hump: Leveraging Multi-Year Site-Specific Impact Evaluation to Derive C&I Lighting Parameters." International Energy Program Evaluation Conference.

TecMarket Works. (June 2004). *California Evaluation Framework.* Sacramento: California Public Utilties Commission. www.calmac.org/publications/California_Evaluation_Framework_June_2004.pdf.

U.S. Environmental Protection Agency (EPA) and U.S. Department of Energy. (July 2006). "National Action Plan for Energy Efficiency." Washington, DC. www.epa.gov/cleanenergy/documents/suca/napee_report.pdf.

York, D.; Kushler, M.; Witte, P. (2007). *Examining the Peak Demand Impacts of Energy Efficiency: A Review of Program Experience and Industry Practice.* Washington, DC: American Council for an Energy-Efficient Economy.

Chapter 11: Sample Design Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

M. Sami Khawaja, Josh Rushton, and Josh Keeling
The Cadmus Group
Portland, Oregon

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68567
September 2017

Contract No. DE-AC36-08GO28308

Chapter 11: Sample Design Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

M. Sami Khawaja, Josh Rushton, and Josh Keeling
The Cadmus Group
Portland, Oregon

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68567
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Suggested Citation

Khawaja, M.S.; Rushton, J.; and Keeling, J. (2017). *Chapter 11: Sample Design Cross-Cutting Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68567. http://www.nrel.gov/docs/fy17osti/68567.pdf

Acronyms

CI	confidence interval
CV	coefficient of variation
DOE	U.S. Department of Energy
ER	error ratio
FPC	finite population correction
GWh	gigawatt hour
HVAC	heating, ventilating, and air conditioning
kWh	kilowatt hour
M&V	measurement and verification
mMcf	million thousand cubic feet
NTG	net-to-gross
SAS	Statistical Analytics Software
SE	standard error
SRS	simple random sampling
UMP	Uniform Methods Project

Table of Contents

List of Figures

List of Tables

1 Introduction

Evaluating an energy efficiency program requires assessing the total energy and demand saved through all of the energy efficiency measures provided by the program. For large programs, the direct assessment of savings for each participant would be cost-prohibitive. Even if a program is small enough that a full census could be managed, such an undertaking would almost always be an inefficient use of evaluation resources.

A cost-effective alternative is to directly assess energy savings for a sample of the program population. However, when a study is based on a random sample rather than a full census, the outcomes of the study are influenced by the particular sample selected for direct evaluation. This random influence is called sampling error. Sampling error introduces an element of uncertainty to every sample-based estimate.

Determining reasonable estimates for quantities of interest is usually a straightforward arithmetic exercise, but quantifying the uncertainty behind such estimates is far more challenging. This document describes the broad principles that apply to all sample-based studies, and it provides specific guidance for applying the procedures most commonly needed in energy efficiency evaluations.

A significant challenge in energy efficiency evaluation is the lack of direct measurement. We can measure energy *consumption*, but energy *savings* is the difference between actual consumption and what consumption *would have been* had energy efficiency measures not been installed. Savings calculations combine consumption measurements with various adjustments to account for technical and behavioral baseline conditions.

Uncertainty can be introduced at every stage of the evaluation, including the sampling, measurement, and adjustment. It is often difficult or impossible to quantify the effect of every potential source of error. Evaluation reports often limit uncertainty discussions to random error (especially sampling error and regression error), because there are well-understood methods for quantifying uncertainty due to random errors. However, a high-quality evaluation should include strategies for mitigating all major sources of uncertainty, and a high-quality report should discuss unquantifiable aspects of uncertainty so research consumers can fully assess the research rigor.

The bulk of this chapter describes methods for minimizing and quantifying sampling error. Measurement error and regression error are discussed in various contexts in other chapters. A broader view of uncertainty is presented in Chapter 12: *Survey Design* and in this chapter's Appendix A.[1]

[1] As discussed in the section "Considering Resource Constraints" of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

1.1 Chapter Organization

The main body of this chapter provides a high-level discussion of the sample design and analysis principles that arise most often in evaluation work. Generally non-technical, this discussion is intended for a wide audience. A more technical, detailed account of important statistical concepts and methods is provided in the appendices.

- Section 2 reviews the statistical terms and concepts routinely encountered in evaluation work.
- Section 3 describes how complex evaluations are broken into components and how component-level research tasks are prioritized.
- Section 4 illustrates the evaluation process through several examples.
- Section 5 discusses validity threats and cost considerations.
- The appendices provide detailed descriptions of the statistical principles and methods that are referenced throughout this document.
    - Section 6: Appendix A discusses general sources and types of errors.
    - Section 7: Appendix B presents fundamental estimates and uncertainty calculations.
    - Section 8: Appendix C presents important sample designs and weighted estimates.

2 Overview

This section presents basic sampling concepts and terminology.

2.1 Sampling and Sample Design

The target group to be studied is called the **population**, and each member of the population is associated with one or more **variables**. The population could be any group of interest, such as program participants, installed measures, or retrofitted sites. A variable can either be a descriptive attribute (such as building type or climate zone) or a numerical quantity (such as square footage, *ex ante* (claimed) savings, *ex post* (evaluated) savings, or air-conditioning tonnage). The primary research objective in a sample-based study is to estimate the population average or total of one or more variables (for example, the total energy and demand savings for all program participants).

Some variables are known through the program database (for example, *claimed* savings) for every member of the population. Other variables (especially *evaluated* savings) can only be obtained through primary data collection and direct estimation. Variables whose values are known for all members of the program population are called **auxiliary**.[2]

A **sample** is a subset of a population selected for direct assessment of one or more variables of interest. The **sample design** describes the exact method by which population members are selected for inclusion in the sample. Sample designs are often informed by auxiliary data such as *claimed* savings estimates or building square footage. **Sample analysis** is the process of estimating population averages or totals and then quantifying the uncertainty in these estimates. The sample analysis may use both sample data and population-level auxiliary data.

Every sample design specifies some element of randomness in the sample selection procedure, but the nature of this randomness varies from one design to the next. Randomization in the sample design forms the basis for calculations that quantify uncertainty in the final estimates, so uncertainty calculations directly depend on the sample design. To yield valid results, the sample analysis must account for the sample design. For example:

- In **simple random sampling** (SRS), each member of the population has probability n/N of being selected,[3] and each individual's inclusion in the sample is unaffected by the particular identities of other members in the sample. If a sample is selected via SRS, then the usual sample mean and standard error formula will yield valid results.
- In **stratified sampling**, auxiliary data are used to partition the population into distinct groups, or strata, and then SRS is performed within each group. In this case, stratum weights are needed to obtain valid analytical results.

[2] In the case of two-phase sampling (Section 8.7), auxiliary data are collected for a large sample through a phone survey or other low-cost interaction. A smaller sample is then selected from the large sample and subjected to intensive measurement and verification. In this case, auxiliary data are known only for the larger sample, but not the entire population.
[3] Here, n is the sample size and N is the population size.

2.2 Uncertainty and Efficiency

Sample design is typically approached with one of two goals:

1. ***To minimize estimator uncertainty, given a fixed amount of study resources.*** In this case, time and budget are the primary constraints. For these projects, the goal is to design a sample that generates the most precise estimate within those constraints.
2. ***To minimize the resources needed to reduce uncertainty to some stated level.*** Often, the evaluation is required to meet a specified confidence-and-precision requirement (typically stipulated by a regulating body or forward-capacity market). In this case, the goal is to minimize time and cost subject to the constraint of meeting this target.

A design is **efficient** if it leads to minimal uncertainty for a fixed research budget. There are many strategies available for designing an efficient study. Energy efficiency program evaluations commonly use one or more of these (in various combinations):

- SRS
- Stratified sampling
- Cluster/multi-stage sampling.

The final design should always be selected to minimize estimation error in light of all available information—including both what is learned through sampling and what is known in advance through auxiliary data. For example, when participant-level *claimed* saving estimates are available, the sample design and analysis plan should use this information to increase efficiency (typically through stratification and/or ratio estimation).

An **estimator** is the particular function (mathematical expression or equation) through which sample data are used to estimate a population quantity. In general, an estimate will not precisely equal its target (for example, the sample mean is unlikely to equal the population mean exactly). The difference between the two—the **sampling error—**can be statistically estimated and, to some degree, controlled through sample design.

Descriptive estimators—such as the mean and standard deviation—can be calculated for any data set. The **mean** is the arithmetic average of the values, while the **standard deviation** is a measure of the variability among observations in the data. In normally distributed data, about 68% of observations are within one standard deviation of the mean, and 95% are within two standard deviations. (Note that a large standard deviation indicates greater dispersion of individual observations about the mean.)

As previously mentioned, the exact value of an estimate depends on the particular sample drawn. Thus, if an entire evaluation were repeated multiple times with a different sample drawn each time, a different estimated value would result for each evaluation.

An estimator is **unbiased** if it tends to be centered at its target quantity. This means that if the entire evaluation (selecting a sample and calculating the estimate based on the sample) were repeated many times, the average of the resulting values would be very near the target population value. The **standard error** (SE) of an estimator quantifies the dispersion that would be observed among these values.[4] The distinction between the standard deviation and the standard error is important. The standard deviation describes variability of the data, while the standard error describes variability of the estimator (for instance, the variability of the sample means obtained from repeated sampling).

For example, in measuring the capacity of a sample of 100 heating, ventilating, and air-conditioning (HVAC) units, the standard deviation for this sample was found to be 25% of the value of the mean capacity. Assuming a normal distribution, approximately 95% of HVAC units in the population should have a capacity within ±50% of the sample mean. However, the standard error is 2.5% of the sample mean ($25\%/\sqrt{100}$). Thus, if we drew repeated samples of 100 HVAC units, the sample means would be within 2.5% of the population mean approximately 95% of the time.

2.3 Confidence and Precision

When data are collected via SRS, the standard error of the sample mean equals the standard deviation of the data, divided by the square root of the sample size.[5] In general, the standard error increases as the standard deviation of the underlying data increases or the sample size decreases.

Statistical methods are available for calculating standard errors for a wide range of estimators. Once an estimator's standard error is known, it is a simple matter to express the estimator's uncertainty through, for example, a **confidence interval** (CI). A CI is a range of values that is believed—with some stated level of confidence—to contain the true population quantity. The **confidence level** is the probability that the interval actually contains the target quantity.

Precision provides convenient shorthand for expressing the interval believed to contain the estimator (for example, if the estimate is 530 kilowatt-hours [kWh], and the relative precision level is 10%, then the interval is 530 ±53 kWh).[6] In reporting estimates from a sample, it is essential to provide both the precision and its corresponding confidence level (typically 90% for energy efficiency evaluations).

For a given data set, an estimate's uncertainty can be expressed in precision terms at any level of confidence. To have higher confidence, it is necessary to take a wider interval, which results in less precision. In other words, when all else is held constant, there is a tradeoff between precision and confidence.[7] As a result, any statement of precision without a corresponding confidence level is incomplete and impossible to interpret. For example, assume the average savings among participants in an ENERGY STAR appliance program is estimated as 1,000 kWh per year, and

---

4 This can be thought of as the standard deviation of the estimator itself, and it may account for multiple sources of random error, including sampling error.

5 This formulation ignores the finite population correction (FPC) (see "Sample Means with FPC" in Appendix C).

6 Note the counterintuitive implication of this standard definition. Low-precision values correspond to narrow intervals and, hence, describe tight estimates. This can lead to confusion when estimates are described as having "low precision."

7 Although there is a close relationship between confidence and precision, these terms are not direct complements of each other. If the confidence level is 90%, there is no reason that the precision needs to be 10%. It is just as logical to talk about 90/05 confidence and precision as 90/10.

the analyst determines this estimate to have 16% relative precision at the 90% confidence level. The same data set and the same formulas may be used to estimate 10% relative precision at the 70% confidence level. If the confidence level is not reported, the second formulation would appear to have less uncertainty when, in reality, the two are identical.

The estimators commonly used in energy efficiency evaluations generally have sampling errors that are approximately normal in distribution.[8] To calculate the bounds for such an estimator, first multiply the estimator's standard error by a z-value.[9] Then add this product to the estimate itself to obtain the CI upper bound, and subtract the product from the estimate to obtain the lower bound.

Note that the z-value depends only on the confidence level chosen for reporting results. That is, for a given estimate $\hat{x}$, the confidence interval is:[10]

$$\hat{x} - z \cdot \widehat{\mathrm{SE}}(\hat{x}) \leq x \leq \hat{x} + z \cdot \widehat{\mathrm{SE}}(\hat{x})$$

In this equation, a z-value of 1.645 is used for the 90% confidence level and a value of 1.960 is used for the 95% confidence level. (These values are tabulated in most statistics textbooks and can be calculated with a spreadsheet.) The absolute and relative precision at the selected confidence level is estimated as:

$$\text{Absolute Precision } (\hat{x}) \; = \; z \cdot \widehat{\mathrm{SE}}(\hat{x})$$

$$\text{Relative Precision } (\hat{x}) \; = \; \frac{z \cdot \widehat{\mathrm{SE}}(\hat{x})}{\hat{x}}$$

The standard error always has the same physical units as the estimator, so absolute precision always has the same physical units as the estimation target. Relative precision, however, is always unit-free and expressed as a percentage.[11]

---

8 This means that if the entire evaluation (drawing a sample and calculating the estimator from the sample) were repeated many times, the resulting estimator values would roughly follow a normal distribution.

9 If the sample size, n, is small, a t-value with n-1 degrees of freedom is more appropriate than a z-value, as z-values will lead to an overstatement of achieved precision. At the 90% confidence level, the choice of t- versus z-value makes little difference for sample sizes greater than 30. The TINV() function in Microsoft Excel can be used to calculate t-values.

10 We have added a "hat" to the SE in this expression. This is to emphasize that any real-life CI would have to rely on a sample-based estimate of the standard error, because the true standard deviation of an estimator cannot be known without perfect knowledge of the population. Inferential statistics in practice substitutes the standard deviation of the sample for the standard deviation of the population. The uncertainty associated with this substitution is treated as negligible. This treatment is usually appropriate, but at very small sample sizes the uncertainties associated with this substitution may become more significant.

Also, strict notational correctness would require a lower case "se" in this equation instead of the "$\widehat{\mathrm{SE}}$." We appreciate the distinction, but do not believe that the failure to distinguish between a function and its generic instance will lead to any errors in practice.

11 Absolute precision is most frequently applied when estimating quantities such as population proportions, which are themselves percentages. In such cases, the expression "… has 5% precision" is ambiguous. It is better to say

Example 1-1

If a program's average savings are estimated as 10.31 kWh and the standard error is calculated as 1.70 kWh, then we have 90% confidence that the true population mean lies within the interval:

$$\begin{aligned} 10.31\text{ kWh} - 1.645 \cdot 1.70\text{ kWh} \quad &\leq \quad \text{average savings} \\ &\leq \quad 10.31\text{ kWh} + 1.645 \cdot 1.70\text{ kWh} \end{aligned}$$

And the precision formulas are

$$\text{Absolute Precision } (\hat{x}) \;=\; 1.645 \cdot 1.70\text{ kWh} \;=\; 2.80\text{ kWh}$$

$$\text{Relative Precision } (\hat{x}) \;=\; \frac{2.80\text{ kWh}}{10.31\text{ kWh}} \;=\; 27.2\%$$

In other words, based on the selected sample, the best estimate of the true (unobserved) population mean is the sample mean (10.31 kWh). We are 90% confident that the true value is within 2.80 kWh or 27.2% of this estimate.

[End of Example]

If the estimated outcome is large relative to its standard error, the estimator will tend to have a small relative precision value at a given confidence level. (Small precision values are desirable.) However, if the amount of variability is large relative to the estimated outcome, the precision will be poor. For example, if the observed average savings is 1,000 kWh and the associated relative precision (at, say, 90% confidence) is 150%, then we are 90% confident that the true average savings is somewhere between negative 500 kWh (which means that the measure actually caused consumption to increase) and 2,500 kWh.

---

either "…has 5% absolute precision" or "… is precise to within five percentage points." (See *Estimating Population Proportions* in Appendix B.)

3 Complex Evaluations: Designing for Multiple Objectives

This section describes sample design and analysis procedures for the research tasks most commonly encountered in energy efficiency evaluations. Evaluations vary in size and complexity. The scope of a given study can be:

- A single program, encompassing several distinct measure groups
- A full portfolio, spanning multiple programs and sectors
- Some collection of measure groups of particular interest to a client.

In the material that follows, the term *study* refers to any of these possibilities. Also, this material mentions—but does not thoroughly discuss—several important statistical concepts; however, these are discussed in detail in *Appendix B. Fundamental Estimates and Uncertainty Calculations* and *Appendix C. Sample Design and Weighted Estimates*.

Most energy efficiency portfolios support a wide range of measures and serve multiple sectors. Complex portfolio evaluations generally include multiple precision requirements at different levels of aggregation. For example, a single evaluation may need to satisfy each of the following:

- Estimate savings to within 10% at the 90% confidence level for each sector (residential, commercial, government/nonprofit, industrial)
- Estimate savings to within 10% at the 90% confidence level for all nonresidential lighting projects combined
- Estimate savings to within 20% at the 90% confidence level for each program in the portfolio.

It would not be difficult to design an efficient study that meets any one of these requirements, but it is much more challenging to design an efficient study that meets all of the requirements simultaneously.

To design an efficient study, the researcher usually engages in some back-and-forth between high-level evaluation requirements and component-level study design details. In all cases, the study design must:

- Lead to valid and essentially unbiased estimates of the object(s) of study
- Meet prescribed confidence and precision targets through valid means
- Be cost-efficient.

The following general steps describe a simplified approach to sample design that relies—to some degree—on trial and error. This approach will lead to an effective and efficient research design for most evaluations. Section 4: *Worked Examples* provides examples illustrating the essential steps, and Appendices A and B give further examples and detailed technical guidance.

1. ***Describe the portfolio structure and the requirements for confidence and precision.*** A complex study may span multiple programs that cover different sectors and

technology groups (for example, custom versus prescriptive). Also, evaluators may be required to provide savings estimates at the study, sector, program, and measure levels.

Often the confidence and precision requirements are imposed through a regulatory process or forward capacity market standard. These values are most commonly set at 90% confidence and 10% precision at the portfolio or sector level, but requirements vary. The evaluator needs to understand which confidence and precision requirements apply to which levels. (That is, at what level—measure, program, sector, portfolio—are savings to be estimated with the stated confidence and precision?) In addition to regulatory precision requirements, clients often require disaggregated results at other levels of precision. A population segment for which an estimate must be reported is called a **reporting domain**.

2. ***Identify the basic sampling and analysis domains.*** At the highest level, the sampling groups usually reflect the structure of the reporting domains. For example, if sector-level savings need to be reported, then residential sampling and analysis will normally be independent of industrial and commercial evaluation activities.[12]

   The basic groups for sampling and analysis are called **domains of study**. There can be multiple evaluation tasks within a study domain. For example, HVAC and lighting savings both need to be evaluated within the commercial sector, but because these measures interact, their evaluation tasks may not be independent. However, each domain's analysis is essentially self-contained and independent of other domains. In the remaining steps, we assume the reporting domains are the same as the domains of study.[13]

3. ***Determine the appropriate stratification.*** The sample sizes and associated data collection costs are directly related to the amount of variability (usually measured with a coefficient of variation or error ratio) in the population. If unit-level savings vary greatly between domain subgroups (for example, measure groups or building types), divide the domain into more homogeneous subgroups (strata). This is called **stratification**. Stratification reduces the sample size needed to obtain a given domain-level precision. (It also allows the evaluator to ensure representation among various subgroups.)

   For example, if domains correspond to sectors, the commercial domain may include the following strata:

Small Retail Lighting	Medium Retail Lighting	Large Retail Lighting
Office Lighting	Office HVAC	Office Plug Load
Small Retail HVAC	Large Retail HVAC	Grocery Refrigeration
Grocery Lighting		

[12] There are exceptions. In some cases, the basic sampling/analysis groups cut across reporting domains, as when sampling and analysis are performed independently within sector-pooled technology groups.

[13] The general principles provided in the appendices remain valid for alternative approaches, but we do not provide step-by-step guidance for all possible approaches.

4. ***Determine the data requirements and estimation strategies within each domain.*** For each group (for example, prescriptive commercial program) or subgroup (for example, offices), use the program database to identify important measure categories (for example, lighting). Then, for each measure category, determine estimation procedures and data needs based on the prevailing measurement and verification (M&V) protocol.

5. ***Record* claimed *contribution,* ex post *uncertainty, and M&V costs for each stratum.*** For each stratum within a domain, determine total *claimed* savings. Based on the M&V protocols (Step 4), note the approximate evaluation cost-per-sample-unit within each measure category. When possible, also include an estimate of the uncertainty parameter (CV or error ratio [ER]) within each category.[14] Measures contributing significantly to total savings and exhibiting significant variability will receive highest levels of evaluation resources.[15] This will reduce the standard error and improve confidence intervals.

6. ***Estimate sample sizes within each domain.*** In the most straightforward cases, the previous step will yield reliable cost, uncertainty, and claimed total estimates. In such a case, implement the cost-weighted Neyman formula (*Appendix C. Sample Design and Weighted Estimates*) to obtain the domain's optimal sample allocation as a function of total sample size n. Adjust n to obtain an efficient domain-level sample allocation, which should meet the precision requirement.

   Sometimes there may be insufficient basis for estimating variation or the reporting requirements may be too complicated to permit a straightforward Neyman allocation. In such cases, the planning process may be simplified by prioritizing measure categories with high *claimed* totals and high uncertainty. The evaluator can then assign initial planning targets of, say, of 10% precision with 90% confidence for each high-priority category. For categories that are not high priority, choose more liberal targets (for instance, 90/20). (These targets may be revised in Step 7.) Sample sizes are then calculated using the formulas provided in *Appendix C. Sample Design and Weighted Estimates*.

7. ***Aggregate Precision to Reporting Requirement Level***. For each reporting level (such as the sector- and study-levels), calculate the expected precision based on the sample allocations obtained in Step 6. If the expected precision at some level falls short of its target, increase the sample sizes in lower-level groups until all precision expectations meet their targets.

   This step is difficult to optimize through a simple formula, but if the calculations in the previous step have been automated, then a gradient-descent algorithm may be used to identify categories that yield the greatest impact on higher-level precision per

[14] This may be based on previous studies' estimates of coefficient of variation. Otherwise, variability may assessed qualitatively (for example, low, medium, or high), based on the evaluator's judgment.
[15] There are, of course, other considerations. See Section 5, *Additional Considerations*, for further discussion.

evaluation dollar and to increase evaluation resources for these categories until higher-level precision estimates meet the evaluation targets.

In cases where a domain's sample allocation is based on evaluator-prioritized precision targets, these targets should be adjusted directly if higher-level precision estimates are significantly higher or lower than the evaluation targets.

8. ***Document the Assumptions and Sampling Plan.*** Document the sampling plan obtained through these steps. Include assumptions about data variability (CVs and ERs) and calculations showing that all precision targets will be met if the observed variability is no greater than what is assumed. At this point, the client and evaluator should agree on the measures to be taken, if any, to adjust sample sizes should early data collection provide evidence that variability assumptions are in error.

Appendix C. Sample Design and Weighted Estimates provides technical guidance about optimizing sample design components. However, the hands-on approach—in which the evaluator prioritizes measure categories and then assigns (and adjusts) precision requirements for each category—is very flexible and sufficient for many applications.

4 Worked Examples

Section 3 described the general procedure for planning a portfolio evaluation at a high level. This section illustrates the basic components of this procedure. The general approach is to begin with lower-level evaluation tasks and then show how these build to a portfolio-level evaluation plan. The discussion makes frequent use of the formulas described in appendices B and C.

4.1 Measure- and Site-Level Evaluation Planning

In most energy efficiency evaluations, populations are segmented by sector: residential, commercial, and industrial.[16] Residential populations tend to be large in number and homogeneous, while the commercial and industrial segments are often smaller and more heterogeneous. Two major considerations drive the sample planning for any measure-level evaluation task:

- The heterogeneity of the relevant population segment (especially with respect to equipment usage patterns)
- The segment's size (in terms of both the number of units in the population and the average savings per unit).

Evaluations in the residential sector often use many different estimators and a variety of data sources. For example, proportions may be estimated from telephone survey data, ratios may be estimated from site visit data, and means may be estimated from end-use metering data. Because residential populations tend to be relatively homogeneous, SRS is the most common sample design in this sector.

Commercial and industrial populations are composed of multiple subsectors (for example, retail, office, grocery, manufacturing, and food processing). Nonresidential portfolios generally offer both prescriptive and custom measures for these sectors. Because the population members vary greatly in size, the expected savings for each measure installation varies from site to site. For example, a convenience store may convert 20 T12 florescent lamps to T8s, but a large office may convert 500 lamps. A well-maintained program database, which would include site-level *claimed* savings estimates, is critical to the efficient evaluation of nonresidential savings. Stratified ratio estimation is a central evaluation tool for these sectors.

4.1.1 Telephone Surveys

Telephone surveys are one of the most common methods of primary data collection in residential evaluations. These surveys are rich sources of data from which a number of population characteristics may be estimated, such as attitudes and opinions, purchasing behaviors, and demographics. Most of the data collected are categorical and are used to estimate proportions (such as the proportion of customers satisfied with the program, or the proportion of customers who actually installed a measure recorded in the program database).

[16] This list is not exhaustive. Other possible segments include: low-income, agricultural, public/institutional, and transportation.

For attitudinal, demographic, and other questions used to inform process evaluation, the uncertainty of a proportion estimate is usually described in terms of absolute precision (see *Appendix B. Fundamental Estimates and Uncertainty Calculations*). Write $e_{\text{abs.}}$ for the absolute precision level. Then the sample size needed to achieve this degree of precision is calculated as:

$$n = \left(\frac{z}{e_{\text{abs.}}}\right)^2 \cdot p(1-p)$$

Here, z is the z-value for the corresponding level of confidence, and p is the true population proportion. The expression $p(1-p)$ obtains its maximum when $p = 0.5$, so an n computed with this value will obtain the desired precision in all cases.

Example 4-1

For part of a process evaluation of a residential energy-education program, a participant survey is used to estimate the proportion of participants who changed their thermostat setting due to the program. The utility wants the survey-based estimate to be within five percentage points (absolute) of the true population proportion, with 90% confidence. If we have no *a priori* knowledge of the true proportion, we use the value with $p = 0.5$ to plan our survey. Then the sample size is:

$$n = \left(\frac{1.645 \cdot 0.5}{0.05}\right)^2 \approx 270.6$$

Thus, a survey sample of 271 participants is needed to ensure the desired level of confidence and precision.

[End of Example]

Note that the finite population correction (FPC) is not used in this formula. The FPC is typically negligible in the residential sector, as program populations tend to be quite large compared to evaluation survey samples.

Telephone surveys may also be used for impact evaluation, but this application should be limited to measures for which:

- No special training is needed to specify the measure and determine that it is installed correctly (For example, energy-efficient showerheads and compact fluorescent lamps satisfy this requirement, but attic insulation does not, because a homeowner may not know the effective R-value of insulation and may not be able to assess installation quality.)
- Average measure savings is well known through other resources.

When these conditions are satisfied, the only information needed to estimate total measure savings is the number of measures installed, and this quantity can be estimated with phone survey data.

When survey-level results are being reported for an impact evaluation, the uncertainty of a proportion estimate is often reported in terms of relative precision. Write $e_{\text{rel.}}$ for the target relative precision level. Then the sample size needed to achieve this degree of precision is calculated as:

$$n = \left(\frac{z}{e_{\text{rel.}}}\right)^2 \cdot \frac{1-p}{p}$$

The expression $(1-p)/p$ does not have any maximum; it increases without bound as p decreases to zero. Thus, some *a priori* lower bound on plausible values for p is needed to calculate the necessary sample size.

If savings at the measure level are not directly reported, but are instead rolled into estimated savings at a higher level for reporting, then measure-level savings is treated as a stratum within the higher level for sample planning.

Example 4-2

Continuing the energy-education example, assume that (1) the results of the participant survey will be used to inform an impact evaluation and (2) average savings among individuals who adjust their thermostats is known through a previous study. Then to estimate program savings, estimate the proportion of participants who adjusted their thermostats.

Consider two possible circumstances:

a. The utility wants the survey-based estimate to be within 20% (relative) of the true population proportion, with 90% confidence. Based on an informal internal evaluation, the utility is confident that at least 40% of the participants have adjusted their thermostats.

   Using

$$\frac{\boldsymbol{(1-p)}}{\boldsymbol{p}} \geq \frac{\boldsymbol{(1-0.4)}}{\boldsymbol{0.4}} = \boldsymbol{1.5},$$

   the sample size is calculated as:

$$n = \left(\frac{1.645}{0.2}\right)^2 \cdot 1.5 \approx 101.5$$

   Thus, a survey sample of 102 participants is needed to ensure the desired level of confidence and precision.

b. The utility does not want results reported at the program level. Instead, estimated program savings are to be rolled into residential sector-level savings for reporting.

   Then this program will be treated as a stratum within the residential domain. Its sample size will be determined through a cost-weighted Neyman allocation applied to the residential sector.

   For this, we will need to record the number of program participants (*N*), the marginal cost of surveying a single participant (*c*), the average savings among participants who adjust their thermostats (*X*), and an *a priori* estimate of the proportion of participants who adjust their thermostats (p_0).

The unit-level standard deviation used in the Neyman allocation is this:

$$s = X \cdot \sqrt{p_0 \cdot (1 - p_0)}$$

This stratum's share of the residential sample will be proportional to $N \cdot s/\sqrt{c}$.

[End of Example]

4.1.2 Verification Site Visits

Verification site visits can be conducted for parameters that are not easily measured by telephone surveys. Common examples are:

- Installation rates (for example, proportion of program-provided CFLs installed)
- Measure Coverage (for example, percent of insulation installed)
- End-use parameters (for example, efficiency rating or thermostat set point).

4.1.2.1 Installation Rates

If there is only one measure per household—as is often the case with water heat, HVAC, and certain appliance measures—then the estimate is a sample proportion, which is analyzed as illustrated in examples 4-1 and 4-2. Note, however, that the marginal cost of a site visit is higher than that of a phone survey, so all else being equal, measures requiring on-site verification will receive smaller shares of the domain-level sample than those requiring only phone surveys. Savings for measures that can have multiple installations at each household or that have measures that vary greatly between sites should be estimated using a mean- or ratio-based method.

Example 4-3

For the evaluation of a direct-mail program that sent three CFLs to each residence within a utility's service territory, assume that the average hours of use and average wattage of replaced lamps are reliably known through a previous study. Write *X* for the product of the average hours of use and the average difference between replaced lamps and program lamps.

Then the research focus is on estimating the number of program bulbs that have been installed. Each residence may have installed 0, 1, 2, or 3 program bulbs (or more if some customers give unwanted CFLs to friends or neighbors). A visited site's savings is estimated as *X* times the number of program bulbs installed at the site. Estimate the average number of installed program bulbs as a simple mean.

To plan this evaluation task, information is used from an earlier evaluation that found the number of program lamps installed at a site was 2.1 on average, with a standard deviation of 1.3.

Consider two possible circumstances:

a. The utility wants the total program savings to be estimated to within 20% (relative precision), with 90% confidence.

Using CV $= \mathbf{1.3/2.1 = 0.62}$, the sample size is calculated as:

$$n = \left(\frac{1.645}{0.2}\right)^2 \cdot (0.62)^2 \approx 25.9$$

Thus, a survey sample of 26 participants is needed to meet the precision target at the stated confidence level.

b. The utility does not want results reported at the program level. Instead, estimated program savings are to be rolled into residential sector-level savings for reporting.

Thus, the program will be treated as a stratum within the residential domain, and its sample size will be determined through a cost-weighted Neyman allocation applied to the residential sector.

For this, record the number of program participants (N), the marginal cost of visiting a single participant (c), the average savings per installed CFL (X), and the *a priori* estimate of the standard deviation of the number of installed lamps per residence (from the previous report, this is 1.3).

The unit-level standard deviation used in the Neyman allocation is $\boldsymbol{s = X \cdot 1.3}$, and the stratum's share of the residential sample should be proportional to $\boldsymbol{N \cdot s/\sqrt{c}}$.

[End of Example]

4.1.2.2 Measure Coverage

Some site visits are made to estimate the proportion of reported savings measures that were actually installed—for example, the proportion of rebated CFLs installed in a home, or the quality and quantity of installed attic insulation. In these cases, the estimation strategy is based on a ratio estimator rather than a proportion- or mean-based estimator (see *Appendix B. Fundamental Estimates and Uncertainty Calculations*).

When measure-level savings must be estimated with a prescribed level of precision and confidence, the sample size formula for the ratio estimator is:

$$n = \left(\frac{z}{e_{\text{rel.}}}\right)^2 \left(\frac{s^{(\text{ratio})}}{\bar{y}}\right)^2$$

Here, $e_{\text{rel.}}$ refers to relative precision and $s^{(\text{ratio})}$ is similar to the standard deviation, but it only captures deviations between *ex post* savings (y_i) and realization-rate-adjusted claimed savings (see *Appendix B. Fundamental Estimates and Uncertainty Calculations*).

When there is no measure-level precision target, the measure is treated as a stratum within sector-level savings. In this case, the measure's share of the sector-level sample should be proportional to

$$N \cdot s^{(\text{ratio})}/\sqrt{c}$$

Where N is the number of participants in the stratum, c is the marginal cost of collecting data for a single participant, and $s^{(\text{ratio})}$ is as above.

Example 4-4

A weatherization program rebates material costs for attic insulation. The program database records the R-value and quantity of rebated insulation for each participant and calculates participant-level *claimed* savings estimates from these data.

To evaluate the program, technicians will visit a sample of participating sites and record the effective R-value (taking into account both the nominal R-value and the installation quality) and the installed quantity. Based on the data collected, *ex post* savings will be estimated for each site, and program savings will be estimated using a ratio-based realization rate. Write x_i for the *claimed* savings of the i^{th} visited site and write y_i for the *ex post* savings. Then

$$\text{Realization Rate} = \frac{\sum_{\text{sample}} y_i}{\sum_{\text{sample}} x_i}$$

The total savings estimate is the realization rate multiplied by the population total of the *claimed* savings values.

In this example, the evaluator is planning the current study using results from the previous year's evaluation. The previous evaluation estimated a realization rate of 75% from a sample of 100 participants. This estimate achieved a relative precision of ±8% with 90% confidence.

Calculate the error ratio, $\text{ER} = s^{(\text{ratio})}/\bar{y}$, based on the values given in last year's report:

$$\frac{s^{(\text{ratio})}}{\bar{y}} = \frac{\sqrt{n} \cdot e_{\text{rel.}}}{z} = \frac{\sqrt{100} \cdot 8\%}{1.645} \approx 0.49$$

Consider two possible circumstances:

a. Program-level results are to be estimated to within 20% (relative precision), with 90% confidence. The sample size is then:

$$n = \left(\frac{z}{e_{\text{rel.}}}\right)^2 (\text{ER})^2 = \left(\frac{1.645}{0.20}\right)^2 (0.49)^2 \approx 16.2$$

Therefore, the evaluator should plan to visit 17 participants to meet the 90/20 target for the realization rate. Because total savings is estimated as the realization rate multiplied by the claimed total, the total savings has the same relative precision as the realization rate.

b. The utility does not want results reported at the program level. Instead, estimated program savings are to be rolled into the sector-level saving estimates for reporting.

Because the program will be treated as a stratum within the residential domain, its sample size will be determined through a cost-weighted Neyman allocation. For this, record the number of program participants (*N*), the marginal cost of visiting a single participant (*c*), and the *a priori* estimate of the standard deviation of the quantity $s^{(\text{ratio})}$.

The stratum's share of the sector sample will be proportional to $\boldsymbol{N} \cdot \boldsymbol{s}^{(\boldsymbol{ratio})} / \sqrt{\boldsymbol{c}}$.

[End of Example]

4.1.2.3 End-Use Parameters

In some cases, the purpose of a site visit is to estimate the value of some end-use parameter, such as the number of linear feet of pipe wrap installed or the technical specifications of an HVAC system. If the program database contains participant-level ex ante information, then total measure savings should be estimated using a ratio estimator. Otherwise, the estimates must be based on the sample mean. In both cases, sample planning for the measure-level evaluation task proceeds as illustrated in the previous examples.

Example 4-5

A site visit is required to estimate the heating capacity of ductless mini-split installed air conditioners (AC) for which customers will receive (or have received) rebates from a residential HVAC program. Unlike the previous residential examples, this program is relatively small, having only 200 participants.

As this is the first evaluation of this program, there is no prior information on the target population. However, the regional technical resource manual refers to a metering study that determined the cooling capacity had a standard deviation of 5.4 kBtu/h. The program implementer assumed that the average mini-split installed AC had a capacity of 18 kBtu/h. Thus, the best estimate of the CV is this:

$$\text{CV} = \frac{s}{\bar{x}} = \frac{5.4}{18} = 0.3$$

To achieve measure-level results having 90% confidence and ±10% relative precision, calculate the initial and, subsequently, the final sample sizes (with finite population correction) as:

$$n_0 = \left(\frac{1.645}{0.10}\right)^2 \cdot (0.3)^2 \approx 24.4$$

$$n = \frac{24.4 \cdot 200}{24.4 + 200} \approx 21.7$$

Thus, visit 22 households to achieve the desired level of precision.

[End of Example]

4.1.3 End-Use Metering

In most cases, end-use metering data are used to estimate some site-specific parameter, such as the average daily hours of use or the average kilowatt (kW) draw. Meter-based estimates are then used to evaluate *evaluated* savings for each metered measure installation. Sampling for end-use metering proceeds as outlined above, with ratio-based estimates used when there is meaningful ex ante information, and mean-based estimates used when no such information is available.

4.2 Domain-Level Evaluation Planning

Sample plans for various levels of reporting domains can be developed after measure-level evaluation tasks have been analyzed and documented, as above. These plans may be based purely on optimization calculations, or they may involve a more hands-on approach (see Step 7 in Section 3).

Example 4-6

For a commercial and industrial (C&I) custom program evaluation, the distribution of participants is shown in Table 1.

Table 1. Example C&I Program Details

Subsector	Participants	End Uses	Percent of Ex Ante Savings
Retail	80	Lighting	25%
Office	65	Lighting, HVAC, Appliances	21%
Restaurant	30	Lighting, Appliances	9%
School	13	Lighting, HVAC	12%
Light Manufacturing	11	Lighting, Motors	33%
Total	**199**	**Lighting, HVAC, Appliances, Motors**	**100%**

To estimate satisfaction with a lighting measure, the evaluator chose to draw a stratified sample. This sample needed to provide a program-level estimate with 10% absolute precision, at the 90% confidence level. Thus, the first step is to determine the overall sample size needed (which is done in the same way as an SRS is determined for a proportion).

$$n_0 = \left(\frac{1.645 \cdot 0.5}{0.10}\right)^2 \approx 67.7$$

$$n = \frac{67.7 \cdot 199}{67.7 + 199} \approx 50.5$$

The results show that calling a total of 51 businesses will achieve the desired level of precision.

To determine how to distribute the sample, use the Neyman allocation, assuming that the variation is proportional to savings. The subsector sample sizes are then calculated as:

$$n_{\text{retail}} = 50.5 \cdot \left(\frac{25\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 12.6$$

$$n_{\text{office}} = 50.5 \cdot \left(\frac{21\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 10.6$$

$$n_{\text{rest.}} = 50.5 \cdot \left(\frac{9\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 4.5$$

$$n_{\text{school}} = 50.5 \cdot \left(\frac{12\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 6.1$$

$$n_{\text{light mfg.}} = 50.5 \cdot \left(\frac{33\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 16.7$$

After rounding the values up to the nearest integer and accounting for the fact that there are only 11 sites in the light manufacturing sector, the final subsector sample sizes are 13, 11, 5, 7, and 11, for a total 47, which is slightly lower than the original 51.

[End of Example]

Example 4-7

To evaluate total savings for the C&I program described by Table 1, regulatory requirements stipulate that results must be within 10% relative precision at the 90% confidence level. Previous experience has shown that, typically, the overall realization rate is approximately 90%, with an ER of approximately 0.4, so the total sample size for the program is:

$$n_0 = \left(\frac{1.645}{0.1}\right)^2 (0.4)^2 = 43.3$$

$$n = \frac{43.3 \cdot 199}{43.3 + 199} \approx 35.6$$

Thus, the initial plan is to visit 36 sites. As before, distribute the sample using the Neyman allocation. There are no data on subsector-specific ERs or CVs, so assume variation within each sector is proportional to ex ante savings.[17] Then for sector *h*, the share of the sample will be proportional to:

$$\begin{aligned} N_h \cdot \frac{s_h}{\sqrt{c_h}} \quad & \propto \quad N_h \cdot \frac{[\textit{ex ante} \text{ total for stratum } h]/N_h}{\sqrt{c_h}} \\ & = \quad \frac{[\textit{ex ante} \text{ total for stratum } h]}{\sqrt{c_h}} \\ & \propto \quad \frac{[\text{stratum } h\text{'s percent of the } \textit{ex ante} \text{ total}]}{\sqrt{c_h}} \end{aligned}$$

Also, evaluation costs differ among subsectors; engineers estimate the following hours are required to evaluate a site for each subsector:

[17] To be precise, assume that within each stratum, the standard deviation of savings is proportional to the stratum's claimed savings average. (If necessary, stratify by size in addition to building type.) For this reasoning, *standard deviation* can either have the usual definition, *s*, or the ratio version, $s^{(\text{ratio})}$. (See *Appendix C*.)

Table 2. Evaluation Times and *Claimed* Savings by Subsector

Subsector	Hours	Proportion of *Claimed* Savings	
Retail	2	25%	
Office	4	21%	
Restaurant	2	9%	
School	4	12%	
Light Manufacturing	8	33%	

Using these estimates as a proxy for cost, allocate sample sizes to each subsector using the cost-weighted Neyman allocation as follows:

$$n_{\text{retail}} \;=\; 35.6 \cdot \left(\frac{25\%/\sqrt{2}}{25\%/\sqrt{2} + 21\%/\sqrt{4} + 9\%/\sqrt{2} + 12\%/\sqrt{4} + 33\%/\sqrt{8}/} \right) \approx\; 5.2$$

$$n_{\text{office}} \;=\; 35.6 \cdot \left(\frac{21\%/\sqrt{4}}{25\%/\sqrt{2} + 21\%/\sqrt{4} + 9\%/\sqrt{2} + 12\%/\sqrt{4} + 33\%/\sqrt{8}} \right) \approx\; 7.4$$

$$n_{\text{rest.}} \;=\; 35.6 \cdot \left(\frac{9\%/\sqrt{2}}{25\%/\sqrt{2} + 21\%/\sqrt{4} + 9\%/\sqrt{2} + 12\%/\sqrt{4} + 33\%/\sqrt{8}} \right) \approx\; 5.2$$

$$n_{\text{school}} \;=\; 35.6 \cdot \left(\frac{12\%/\sqrt{4}}{25\%/\sqrt{2} + 21\%/\sqrt{4} + 9\%/\sqrt{2} + 12\%/\sqrt{4} + 33\%/\sqrt{8}} \right) \approx\; 7.4$$

$$n_{\text{light mfg.}} = 35.6 \cdot \left(\frac{33\%/\sqrt{8}}{25\%/\sqrt{2} + 21\%/\sqrt{4} + 9\%/\sqrt{2} + 12\%/\sqrt{4} + 33\%/\sqrt{8}} \right) \approx\; 10.4$$

After rounding the values up to the nearest integer, the final subsector sample sizes are 6, 8, 6, 8, and 11, for a total 39. This represents the allocation that optimizes the balance between precision and cost.

4.3 Portfolio-Level Evaluation Planning

This section illustrates the planning process outlined in Section 1.3 through an extended example of an energy efficiency portfolio evaluation. The utility promotes efficiency measures in the residential, institutional (government and nonprofit), commercial, and industrial sectors. Table 3 shows program sizes.

Table 3. Claimed Savings by Sector

Sector	*Claimed* kWh Total
Residential	2,900,000
Institutional	2,200,000
Commercial	3,300,000
Industrial	3,000,000
Total	**11,400,000**

This evaluation entails estimating total savings to within 10% for each sector and to within 5% for the entire portfolio (all precision values assume 90% confidence). Sampling and analysis are to be performed separately within each sector (thus, data collected in the commercial sector has no bearing on estimates related to the industrial sector).

- Steps 1 and 2 are immediate: Report the savings for each of the four sectors, and the sectors are the domains of study.
- For Step 3, stratify each domain by measure group and size.
- For Step 4, examine the program database to determine the specific measures and measure groups that contribute to savings within each sector.

Table 4 shows savings by measure category for the residential program.

Table 4. Residential Program Data

Measure Group	*Claimed* kWh
Lighting	1,800,000
HVAC	600,000
ENERGY STAR Appliances	500,000
Total	**2,900,000**

This utility recently completed a study of ENERGY STAR appliances, so deemed values are considered acceptable for that program, so long as installation rates are directly evaluated. Then telephone surveys will provide acceptable data, and a proportion estimator will be appropriate for estimating savings. Stratification may also be appropriate if there are distinct participant groups for which installation rates may vary.

After reviewing the M&V protocols, the evaluator determines that (1) usage loggers are needed for evaluating savings from lighting measures and (2) interval metering is needed for evaluating HVAC savings. The final verified savings for both measure types will be determined through engineering calculations. After calculating savings for measures in the sample, ratio estimators will be used to evaluate total program savings for both measure groups.

For Step 5, consider the data to be used in the savings calculations to (1) determine average M&V costs for sampled units within each measure category and (2) anticipate variability within each group. (This process was illustrated in Section 4.1: *Measure- and Site-Level Evaluation Planning*.) Then use the cost-optimized allocation formula to determine the sample fraction for each group (Step 6). The results are summarized in Table 5.

Table 5. Cost, Variability, and Sample Fractions for Residential Sector

Measure Group	Evaluation Cost per Unit	Anticipated Variability	Average Claimed kWh	*Claimed* Standard Deviation	Sample Fraction
Lighting	$2,000	0.4 (ER)	200	80	48.3%
HVAC	$2,500	0.6 (ER)	2,400	1,440	21.6%
ES Appliances	$100	0.2 (CV)	250	50	30.0%

In Table 5, variability entries are based on experience with similar evaluation tasks. Average *claimed* values are based on program data and the standard deviations are the products of average savings and the error ratios or coefficients of variation. The sample fractions are calculated using the formula from *Planning and Optimizing Stratified Designs* (Appendix C).

Continuing Step 6, use the standard error formulas to determine the standard error for estimated total savings as a function of sample size. After some experimentation, the evaluator determines a residential sample allocation that should yield the 90/10 target for the sector. In Table 6, measure-level standard errors are based on estimator-specific standard error formulas. The total standard error is the square root of the sum of squared measure-level standard errors.

Table 6. Preliminary Sample Allocation for Residential Sector

Measure Group	Claimed kWh Total	*Claimed* Standard Deviation	Sample Size	Standard Error (Evaluated Total)	Relative Precision
Lighting	1,800,000	80	30	131,453	12.0%
HVAC	600,000	1,440	13	99,846	27.4%
ES Appliances	500,000	50	19	22,942	7.5%
Total	2,900,000	NA	62	166,660	9.5%

Repeat this process for the institutional, commercial, and industrial sectors. This is the more hands-on approach to Step 6, which begins with stipulated group-level precision targets, and usually leads to more back-and-forth iterations. Note that the more technical approach is also valid.

For Step 7, collect sector-level *claimed* savings totals and standard errors and use the formula for the standard error of a sum of independent estimates to estimate the standard error and precision at the portfolio level.

Table 7. High-Level Standard Errors

Sector	*Claimed* kWh Total	Precision	Standard Error
Residential	2,900,000	9.5%	166,660
Institutional	2,200,000	10%	133,739
Commercial	3,300,000	10%	200,608
Industrial	3,000,000	10%	182,371
Total	**11,400,000**	**4.6%**	**318,243**

The implied portfolio-level precision is $1.645 \cdot 318{,}243/11{,}400{,}000 = 4.6\%$, so this sample allocation will meet all precision targets if our CV and ER assumptions hold.

If the estimated precision value had been higher than the target, the evaluator would increase the sample sizes incrementally for the influential sector(s) with the lowest marginal sampling costs until the overall precision was achieved.

[End of Example]

5 Additional Considerations

The following sections discuss important considerations when choosing both a sample size and design.

5.1 Threats to Validity

The fundamental assumption in a design-based sample analysis is that population members have been sampled according to the rules specified in the sampling plan. When factors external to the sample plan affect the final sample, the study's validity may be compromised. In particular, specific external factors may lead to biased estimators and incomplete pictures of uncertainty.

The following are validity threats that commonly arise in impact evaluations.[18]

1. **Non-Coverage.** Validity is threatened when significant population segments are not included in the sample frame. The result is that values calculated from the sample cannot then be said to be representative of the entire population.
2. **Non-Response.** This type of threat occurs in every sample-based study for which population members have the option of refusing to be included. If certain types of households are more likely to refuse to participate or to respond to certain questions, the values calculated from the sample will understate the contribution of this portion of the population.
3. **Self-Selection**. In evaluation activities where participation is voluntary, some groups of people may be more likely to participate than others. This may be associated with demographics, education level, personal attitudes, or any number of unobservable factors. If this is the case, the estimate from these samples may not be completely representative.
4. **Measurement Error.** At times, data collection done either through metering or survey instruments may not be completely accurate.[19] Metering results can be biased by equipment failure, incorrect placement, or poor calibration. Survey instruments are vulnerable to a variety of threats that can be thought of as types of measurement error, such as: construct error, ambiguous wording of questions, and respondent social bias.

5.2 Cost Considerations

There is always a tradeoff between cost and precision. Although some gains in precision can be made through a thoughtful sample design, increasing the sample size always leads to better precision. However, the cost of doing so can be prohibitive.

---

[18] Threats to validity and strategies for mitigating their effects are explored in greater detail in *Appendix A*. For issues specific to survey instruments, see also the "Survey Design and Implementation for Estimating Gross Savings" chapter of this document.

[19] In most metering applications, this measurement error is ignored, particularly when data sources are utility-grade electricity or natural gas meters. However, other types of measurements—such as flow rates in water or air distribution systems—can have significant errors. The magnitude of such errors is often not large enough to warrant concern in a program evaluation and is largely provided by manufacturer's specifications.

The general precision equation can be written in this form:

$$\text{Precision} = \text{confidence level} \sqrt{\frac{\text{variance}}{\text{sample size}}}$$

Precision is a function of three factors: the confidence level (z), variance (s^2), and the sample size (n). The confidence level is fixed for a given study (typically at 90% for energy efficiency evaluations). The population variance does not change with sample size either, so the only factor under the evaluator's control in this equation is the sample size. However, precision is not improved at rate proportional to the sample size, but by the square root of the sample size. This is an important consideration in evaluation planning, as the cost-sample-unit is often linear, while improvements in precision are not.

Example 5-1

In conducting a metering study of commercial lighting to determine average hours of operation, the evaluator first performs a literature review. The effort reveals past studies showing that commercial lighting hours of operation typically vary with a CV of 0.5. When considering costs, the evaluator estimates each site will cost $1,000 for travel, data collection, and analysis. Figure 1 compares cost to precision.

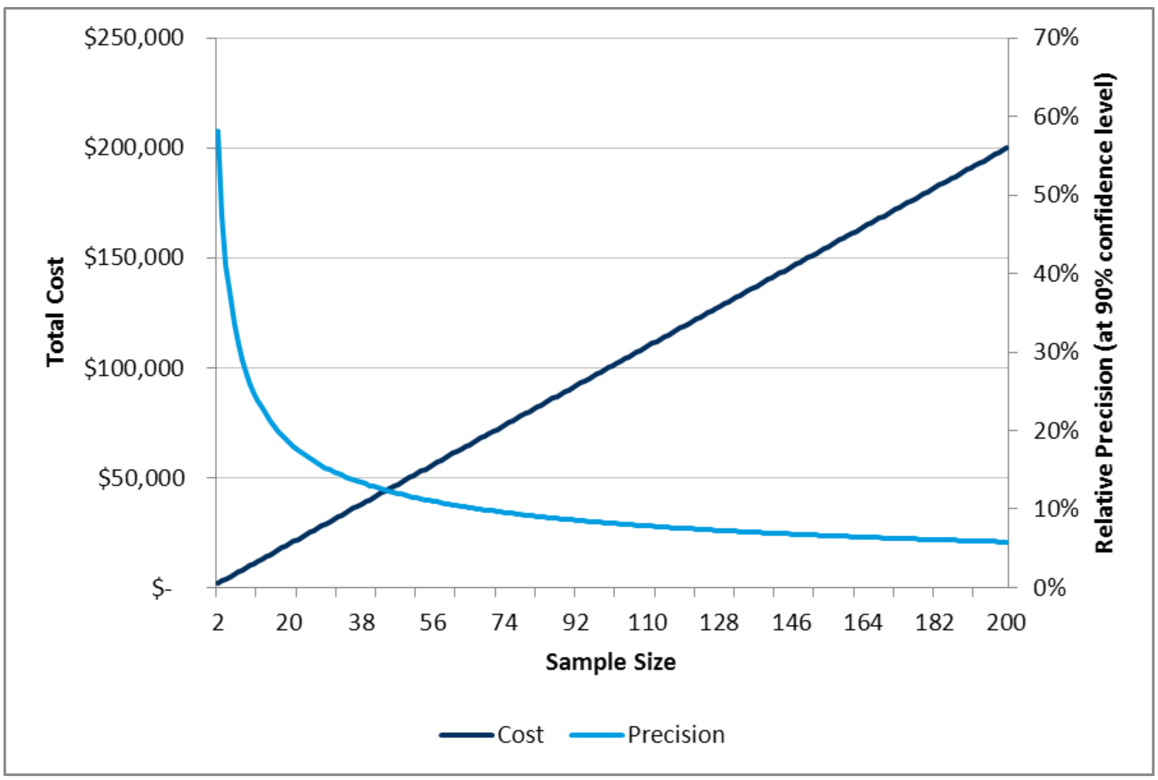


Figure 1. Example: cost vs. precision

So, visiting 70 sites to achieve ±10% relative precision (at the 90% confidence level) will cost $70,000. However, visiting only two sites (the minimum to calculate precision) would result in relative precision of ±58% at a cost of $2,000. Thus, given repeated experiments, a 1% improvement in precision can be expected to cost an average of approximately $1,417.

If the evaluator chose to sample an additional 70 sites, the results would have a relative precision of ±7% at a total cost of $140,000. While the costs doubled, the precision only improved by approximately one third. Thus, average cost for a 1% increase in precision has now ballooned to approximately $23,333.

[End of Example]

5.3 Varying Uncertainty

In some cases, variation in the estimates of interest may differ in magnitude. If these measures are being combined, then the overall uncertainty of the final outcome is a function of those measures with large and small variation. As precision increases with variability (shown in the general equation repeated here), the overall sample will be more efficient when those measures with higher savings variation are allotted larger samples.

$$\text{Precision} = \text{confidence level}\sqrt{\frac{\text{variance}}{\text{sample size}}}$$

It is common practice in energy efficiency evaluations to estimate different parameters of an algorithm by different methods. One parameter may come from a phone survey, another from site visits, and a third may come from a secondary source. It is critical in these evaluations to identify the parameters having the greatest potential impact on overall uncertainty and then target them accordingly.

For example, in an evaluation conducted to estimate the savings of a residential energy-efficient showerhead program, the main inputs are hours of use, flow rate, and the installation rate. While installation rate and hours of use can be measured by phone survey, the flow rate must be measured on site. In this study, the evaluator knows that the CV of hours of use is much higher than the CV of flow rate. Thus, applying a sampling strategy that allots more of the sample to phone surveys and less to site visits could be more efficient than an equal allotment.

5.4 Outcome of Interest

As shown in the preceding example, it is critical to determine the true value of increased precision. Making this determination entails not only cost considerations, but knowing the value to the overall measure of interest. In an energy efficiency evaluation, this is most often total portfolio gross and/or net energy savings. If precision targets are set at the portfolio level, then the relative precision of a portfolio of programs is calculated as follows:

$$\text{Relative Precision of Portfolio} = 1.645 \cdot \left(\frac{1}{\sum_{i=1}^{m} \widehat{\text{savings}_i}}\right) \cdot \sqrt{\sum_{i=1}^{m}\left(SE[\widehat{\text{savings}_i}]\right)^2}$$

This formula follows from results presented in Appendix B. Fundamental Estimates and Uncertainty Calculations.

In Example 4-1, a 3% improvement in precision may justify an additional $70,000 in costs if the savings in this stratum represents a large proportion of total savings. If, however, a given measure makes up only 10% of total program savings, then a 1% improvement in precision at the measure level only contributes approximately 0.1% to the precision at the program level. Thus, both cost and value should be considered when choosing how to allocate resources effectively.

6 Appendix A. Sources and Types of Error

This appendix provides an introduction to how uncertainty is classified in evaluation applications, and it discusses systematic error and random error unrelated to sampling.

6.1 Sources of Uncertainty

As a measure of the "goodness" of an estimate, *uncertainty* refers to the amount or range of doubt surrounding a measured or calculated value. Any report of gross or net program savings, for example, has a halo of uncertainty surrounding the reported relative value to the true values (which are not known). As defined this way, uncertainty is an overall indicator of how well a calculated or measured value represents a true value. Without some measurement of uncertainty, it is impossible to judge an estimate's value as a basis for decision-making.

Program evaluation seeks to estimate energy and demand savings with reasonable accuracy. This objective may be affected by:

- **Systematic error** (that is, not occurring by chance), such as non-coverage, non-response, self-selection, and some types of measurement errors
- **Random error** (that is, occurring by chance), attributable to using a population sample rather than a census to develop the calculated or measured value. This error type can also be the result of some types of measurement error.[20]

The distinction between systematic and random sources of error is important because different procedures are required to identify and mitigate each. Although the amount of random error can typically be estimated using statistical tools, other means are required to estimate the level of systematic error. Because additional investment in the estimation process can lead to reductions in both types of error, tradeoffs between evaluation costs and reductions in uncertainty are inevitably required.

6.2 Sources of Systematic Error

Systematic errors typically occur from the way data are measured, collected, and/or described:

1. **Measured.** At times, equipment used to measure consumption may not be completely accurate. Human errors (for example, errors in recording data) may also cause this type of error. Metering results can be biased by equipment failure, incorrect placement, or poor calibration.[21] Survey instruments are vulnerable to a variety of threats that can be thought of as types of measurement error, such as construct error, ambiguous wording of questions, and respondent social bias.

[20] Note that measurement error may be systematic or random. For example, a meter that is not properly calibrated and consistently under- or overestimates a measurement exhibits systematic error. A meter that is only accurate within a given interval is said to have random error within that interval.

[21] Such errors will bias measurements within a site. However, because the magnitude and direction of the bias may differ from one site to the next, these errors may be viewed as random (not systematic) from the point of view of the broader evaluation, provided the errors are not similar across sites.

Measurement error is reduced by investing in more accurate measurement technology, establishing clear data collection protocols, and reviewing data to confirm they were accurately recorded. In most applications, this error source is ignored, particularly when data sources are utility-grade electricity or natural gas metering equipment. However, other types of measurements can have significant errors.

2. **Collected.** *Non-coverage errors* can occur when some parts of a population are not included in the sample. This can be a problem because the value calculated from the sample will not accurately represent the entire population of interest. Non-coverage error is reduced by investing in a sampling plan that addresses known coverage issues. For example, a survey implemented through several modes (such as phone, Internet, and mail) can sometimes address known coverage issues, assuming that non-coverage is related to the means of communication. However, in some cases there is little to do beyond clearly stating that some hard-to-reach segment of the population was excluded from the study.

   Non-response errors occur when some portion or portions of the population having certain attitudes or behaviors are less likely to provide data than are other population portions. In a load research or metering study, if certain types of households are more likely to refuse to participate—or if researchers are less likely to be able to obtain required data from them—the values calculated from the sample will understate the contribution of this portion of the population and over-represent the contribution of sample portions more likely to respond. In situations where the underrepresented portion of the population has different consumption patterns, non-response error is introduced into the value calculated from the sample. Non-response error is addressed through investments that increase the response rate, such as incentives and multiple contact attempts.

   The converse of non-response errors are *self-selection errors*. In evaluation activities where participation is voluntary, some groups of people may be more likely to participate than others. This may be associated with demographics, education level, personal attitudes, or any number of unobservable factors. If this is the case, the estimate from these samples may not be completely representative. Self-selection bias is best addressed by conducting studies in which participation is mandatory, although this is typically infeasible. Establishing representative quotas by demographics believed to be associated with self-selection may also mitigate these effects.

   Researchers often use "weights" in deriving their final estimates. These weights are means of adjusting the representativeness of the sample to reflect the actual population of interest. For example, if the proportion of single-family respondents is 70% in the sample but is 90% in the population, a weight of 90/70 can be used to increase the representativeness of single-family responses.

3. **Described (modeled).** Estimates are created through statistical models. Some are fairly simple and straightforward (for example, estimating the mean), and others are fairly complicated (for example, estimating response to temperature through regression models). Regardless, modeling errors may occur due to using the wrong model, assuming inappropriate functional forms, including irrelevant information, or

excluding relevant information (for example, in modeling energy use of air conditioners, the evaluator used cooling degree days only). In another example, home square footage or home type may not be available, so the statistical model will attribute all the observed differences in energy use to temperature, although clearly a portion of the use is attributable to the home size. This model will introduce systematic error.

Bias in regression estimates resulting from the omission of a relevant variable is also a well-known phenomenon. While evaluators use experience, economic theory, and engineering principles to prevent this type of bias, there is no statistical procedure to testing for this bias.

Reference manual assumptions are another potential source of modeled error. Technical reference manuals describe estimation procedures that are designed to balance evaluation rigor with practical concerns. Engineering assumptions and stipulated or deemed parameter values can introduce bias.

However, if a deemed value is obtained from a study that reports the value's standard error, then this standard error can be incorporated into a later evaluation, provided the study's target population is similar to the population being evaluated. In this case, the unknown bias can be accounted for within the evaluation's standard error calculations.

6.3 Sources of Random Error

Most random errors are due to sampling, measurement, or regression/extrapolation.

1. **Sampling.** Whenever a sample is selected to represent the population—whether the sample is of appliances, meters, accounts, individuals, households, premises, or organizations—there will be some amount of random sampling error. Any selected sample is only one of a large number of possible samples of the same size and design that could have been drawn from that population. Sampling error and strategies for mitigating it are discussed in detail in the rest of this document.

   The primary topic of this chapter is the mitigation and quantification of sampling error.

2. **Measurement.** In a survey, random measurement error may be introduced by factors such as respondents' incorrectly recalling dates, expenses, or by differences in a respondents' mood or circumstances, which affect how they answer a question. Technical measurements can also be a source of measurement error. (See item 1 and footnote 20 in the systematic error list.)

   These types of random measurement error are generally assumed to even out, so that they do not introduce systematic bias, but only increase the variability. For this reason, researchers often do not attempt to quantify the potential for bias due to random measurement error. However, measurement error can still be a source of variability, and researchers are encouraged to include this source of uncertainty in standard error calculations when it presents a significant threat to validity.[22]

3. **Regression.** Regression error may arise at either the measure/site level, or at the population/stratum level.

   Site-level regression error arises when site-level savings estimates are obtained through regression (where a separate model is fitted to each site's data, and each site's savings is estimated through some function of the fitted parameters). For most site-level regression procedures, standard regression theory will provide a way to estimate the standard error of each site's savings estimate. These standard errors can then be accounted for in an evaluation's uncertainty calculations using methods similar to those applied in two-stage sampling. (See Section 8.7: *Two-Phase (Nested) Sampling*, of Appendix C. Also, ASHRAE Guideline 14 provides further details.)

   Population-level regression error arises when a single regression model is fit to data from multiple sites—possibly the entire population of sites that installed some program measure of interest. For example, a billing analysis may estimate program-wide natural gas savings due to high-efficiency residential furnaces by fitting a regression to billing data from all program participants and a control group of nonparticipants. The standard error of such regression-based estimates can be calculated with standard regression-related methods. Because the standard error applies to the estimate of total savings due to a measure—rather than site-level savings—this standard error is rolled up into sector- or portfolio-level savings uncertainty using the root-sum-of-squared-error formula. (In other words, it is treated in precisely the same manner as stratum-level sampling error.)

6.4 Mitigating Systematic Error

Determining the steps needed to mitigate systematic error is a more complex problem than mitigating random error, because various sources of systematic error are often specific to individual studies and procedures. To mitigate systematic error, evaluators typically need to invest in additional procedures (such as meter calibration, a pretest of measurement or survey protocols, a validation study, or a follow-up study) to obtain additional data to assess differences between participants who provided data and those who did not.

To determine how rigorously and effectively an evaluator has attempted to mitigate sources of systematic error, the following may be examined:

1. Were measurement procedures (such as the use of observational forms or surveys) pretested to determine if sources of measurement error could be corrected before the full-scale fielding?
2. Were validation measures (such as repeated measurements, inter-rater reliability, or additional subsample metering) used to validate measurements?
3. Was the sample frame carefully evaluated to determine what portions of the population, if any, were excluded in the sample? If so, what steps were taken to

[22] ASHRAE Guideline 14-2002 and Guideline 2002R offer extensive guidance on accounting for measurement error. Also, see Section 8.6, *Two-Stage Sampling for Large Projects* in this document for a related discussion.

estimate the impact of excluding this portion of the population from the final results?

4. Were steps taken to minimize the effect of non-response or self-selection in surveys or other data collection efforts? If non-response appears to be an issue, what steps were taken to evaluate the magnitude and direction of potential non-response bias?
5. Has the selection of formulas, models, and adjustments been conceptually justified? Has the evaluator tested the sensitivity of estimates to key assumptions required by the models?
6. Did trained, experienced professionals conduct the work? Was the work checked and verified by a professional other than the one conducting the initial work?

Many evaluation reports do not discuss any forms of uncertainty other than sampling error, which is quantified through confidence intervals for energy or demand savings. This is misleading because it suggests that (1) the confidence interval describes the total of all uncertainty sources (which is incorrect) or (2) the other sources of uncertainty are not important relative to sampling error. Sometimes, however, uncertainty due to other sources of error can be significant. A quality report should discuss all potentially significant sources of uncertainty so that research consumers can fully assess the evaluation's rigor.

6.4.1 Measurement Error

Measurement error can result from inaccurate mechanical devices (such as meters or recorders), inaccurate recording of observations by researchers, or inaccurate responses to questions by study participants. Basic human error occurs in taking physical measurements or conducting analyses, surveys, or documentation activities.

For mechanical devices—such as meters or recorders—it is theoretically possible to perform tests with multiple meters or recorders of the same make and model to assess the variability in measuring the same value. However, for meters and most devices regularly used in energy efficiency evaluations, it is more practical to use manufacturer or industry study information on the likely amount of error for any single piece of equipment.

Assessing the level of measurement error for data obtained from researchers' observations or respondents' reports is usually a subjective exercise, based on a qualitative analysis. This is because it is often impossible to make objective quantitative measures of these processes. The design of recording forms or questionnaires, the training and assessment of observers and interviewers, and the process of collecting data from study participants are all difficult to quantify.

Special studies of a subsample can be used to provide an assessment of the uncertainty potential in evaluation study results. For example:

- It is possible to have more than one researcher rate the same set of objects to evaluate the level of agreement between ratings.
- By conducting short-term metering of specific appliances for a subsample, an evaluator can verify information about appliance use.

- Participants can be re-interviewed to test their answers to the same question at different times.
- Pretests or debriefing interviews can be conducted with participants to determine how they interpreted specific questions and constructed their responses.

6.4.2 Non-Coverage and Non-Response

Another challenge is estimating the effect of excluding a portion of the population from a sample (sample non-coverage) or of the failure to obtain data from a certain portion of the sample (non-response). The data needed to assess these error sources are typically the same as those needed to resolve the errors; but such data are usually unavailable.

However, for both non-coverage and non-response, it is sometimes possible to design special studies to estimate the uncertainty level introduced.

- If a particular portion of the population was not included in the original sample design, it is possible to conduct a small-scale study on a sample of the excluded group. For example, conducting a special study of respondents who are in a particular geographical area or who are living in a certain type of housing can help determine the magnitude and direction of differences in calculated values for this portion of the population.
- In some situations—such as a survey—it is also possible to conduct a follow-up study of a sample of members from whom data were not obtained. This follow-up would also provide data to determine if non-respondents were different from respondents, as well as an estimate of the magnitude and direction of the difference.

7 Appendix B. Fundamental Estimates and Uncertainty Calculations

This section describes basic estimators commonly used in energy efficiency evaluations. Standard errors and other important formulas are also provided. These are fundamental to quantifying uncertainty, and they provide the foundation for basic sample design. For all formulas and examples in this section assume the data are collected through a simple random sample of size n from a very large population.[23]

Many research questions can be phrased in terms of:

- A population *average*, such as average savings among program participants or proportion of participants with gas heat
- A population *total*, such as total savings among all program participants or total number of customers with gas heat.

For consistency, this section's results are generally expressed in terms of averages. To estimate a population total, simply multiply the estimated average by the population size. The resulting estimate's standard error is the population size times the standard error of the average estimate. Because both the estimator and its standard error are multiplied by the population size, the relative precision is unaffected when translating between estimates of population averages and estimates of population totals.

7.1 Estimating a Population Proportion

Many energy efficiency evaluation tasks use survey data, which are typically used to estimate proportions. To estimate the proportion of the population having characteristic x (such as the proportion of utility customers who are aware of a given program), we use this formula:

$$\hat{p} = \frac{n_x}{n}$$

Where:

n_x = the number of sample points with characteristic x

n = the sample size.

To quantify the uncertainty surrounding this estimate, calculate the standard error and then calculate the precision.

[23] When the population is not very large, a non-negligible finite population correction will apply to standard errors. Simple random samples with finite population corrections are discussed in detail in Section 8.1, *Simple Random Sampling* in Appendix C.

The standard error of a proportion is most often[24] calculated as:

$$\widehat{\mathrm{SE}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The absolute precision is then calculated as:

$$\text{Absolute Precision}(\hat{p}) \; = \; z \cdot \widehat{\mathrm{SE}}(\hat{p})$$

Note that the absolute precision equation does not involve dividing by the original estimate. This is different from energy savings estimates, where uncertainty is generally expressed in terms of relative precision. However, in process-related contexts, relative precision for a proportion can be a confusing measure, as the next example shows.

Example B-1
In a survey of 400 participants regarding their experience with a rebate program, we estimate the proportion of program participants satisfied with their rebate amount as $\hat{p} = 92\%$. We can then calculate the absolute precision at the 90% confidence level:

$$\text{Absolute Precision}(\hat{p}) \; = \; 1.645\sqrt{\frac{0.92(1-0.92)}{400}} \; = \; 2.2\%$$

Thus, we are 90% confident that the proportion of participants satisfied with the rebate is between 89.8% and 94.2%.

The relative precision, however, is calculated as:

$$\text{Relative Precision}(\hat{p}) \; = \; \frac{1.645\sqrt{\frac{0.92(1-0.92)}{400}}}{0.92} \; = \; 2.4\%$$

The relative and absolute formulations are both describing the same range of values, but the relative version expresses the confidence interval (CI) width as a proportion of a proportion. It says the CI has a width of 2.4% of 92%.

Not only is this confusing, it also leads to precision values that depend on how study results are communicated. The same study results could be communicated in terms of the proportion of participants who are *not* satisfied with the rebate amount. In this case, we have:

[24] When $\hat{p}$ is very close to one or zero, confidence intervals should be calculated through alternative means, such as the exact binomial method (see Example B-2). An oft-cited rule is that the exact method should be used if either n_x or $n - n_x$ is less than five.

$$\text{Absolute Precision}(1-\hat{p}) = 1.645\sqrt{\frac{0.08(1-0.08)}{400}} = 2.2\%$$

$$\text{Relative Precision}(1-\hat{p}) = \frac{1.645\sqrt{\frac{0.08(1-0.08)}{400}}}{0.08} = 27.8\%$$

While the absolute precision is the same as before, the relative precision is more than 10 times larger than previously calculated. As a result, someone reading the results might think the "unsatisfied" estimate is less precise than the "satisfied" estimate, despite the fact they convey identical information.

[End of Example]

In general, we recommend that precision for population proportions be expressed in absolute terms, especially when the research question is attitudinal or demographic. However, when the research target is a direct indicator of savings (such as the proportion of program-provided measures that are actually installed), relative precision may be preferred.

In Example B-1, the population proportion was estimated as $\hat{p} = 92\%$. Because the sample was of size $n = 400$, the data must have comprised $n_x = 368$ positive survey responses and $n - n_x = 32$ negative responses. Neither of these is less than five, so we were justified in using methods that assume $\hat{p}$ has an approximately normal sampling error. The next example illustrates the **exact binomial method**, which does not require the normality assumption.[25]

Example B-2

To verify the installation of measures that are recorded in a program database, we survey 50 participants, of whom 48 indicate they have installed the measure noted in the database. Thus, we estimate the percentage of participants who have installed the measure as $\hat{p} = 96\%$. However, with only two negative survey responses, we cannot say that the sampling error of $\hat{p}$ is approximately normal. Therefore, we need a method for obtaining a confidence interval that does not appeal to normality through a *z*-value. One option is the exact binomial method.

In a survey of $n = 50$ randomly selected people, the number of positive responses, n_x, follows a binomial distribution with 50 trials and an unknown "success" probability *p* for each trial. To construct a 90% CI for *p*, we calculate the upper and lower CI bounds separately.

[25] The exact binomial never understates uncertainty, but it often overstates it. This conservatism may be appropriate for some applications, and inappropriate for others. See Agresti (2003) or Brown, Cai, and DasGupta (2001) for details and alternative methods. In spite of the apparent simplicity of estimating a population proportion, there is no full consensus on the most desirable confidence interval for this problem among practicing statisticians. Alan Agresti, Brent Coull, George Casella, and others have attached insightful comments to the Brown, Cai, and DasGupta paper.

For the CI lower bound, we must answer the question, "*What is the smallest p for which the probability of obtaining 48 or more 'successes' is less than 5%?*" In Excel, this question can be answered using

$$\texttt{=Binom.inv}(50, p, 0.95)$$

For a given value of p, this function returns the smallest integer m for which the probability that $n_x \leq m$ is at least as large as 0.95.

If we choose a value p for which the function returns $m = 48 - 1$, then we know that the probability of 48 or more successes is no greater than 5% for the chosen p.

After finding a p for which the function returns a value of 47, we adjust p upward until the function returns a value of 48. Write $\hat{p}_{\text{lower}}$ for the largest p for which the function returns a value of 47. Then we are 95% confident that $p \geq \hat{p}_{\text{lower}}$.

In this example, the exact binomial method yields $\hat{p}_{\text{lower}} = 87.9\%$. A similar process yields the CI upper bound, $\hat{p}_{\text{upper}} = 99.3\%$. Thus, our estimate is $\hat{p} = 96\%$, and the exact binomial 90% confidence interval for p is

$$87.9\% \ \leq \ p \ \leq \ 99.3\%$$

For comparison, the normal-based confidence interval is

$$91.4\% \ \leq \ p \ \leq \ 100\%$$

The normal-based confidence interval understates uncertainty relative to the exact binomial confidence interval.

[End of Example]

In an extreme case, all survey responses may be affirmative. Then with no variability in the data, there is no basis for constructing a normal-based CI. However, it would not be credible to report 100% confidence that 100% of the population is in the affirmative category. The exact binomial method will yield a credible CI in such cases.

7.2 Using a Sample Mean to Estimate a Population Mean

Evaluations often need to estimate the average energy consumption for particular equipment types, such as residential refrigeration. When no useful auxiliary information is available,[26] the population average is estimated by the sample mean,

$$\bar{x} = \frac{\sum x_i}{n}$$

[26] Auxiliary information is discussed in the next section.

To quantify the uncertainty surrounding this estimate, calculate the standard error and then the precision. The sample mean's standard error is:

$$\widehat{\mathrm{SE}}(\bar{x}) = \frac{s}{\sqrt{n}}$$

Here, the sample standard deviation, s, is calculated as:

$$s = \sqrt{\frac{\sum(\bar{x} - x_i)^2}{n-1}}$$

The absolute and relative precision are then calculated as:

$$\text{Absolute Precision}(\bar{x}) \;=\; z \cdot \widehat{\mathrm{SE}}(\bar{x}) \;=\; z \cdot s/\sqrt{n}$$

$$\text{Relative Precision}(\bar{x}) \;=\; z \cdot \frac{\widehat{\mathrm{SE}}(\bar{x})}{\bar{x}} \;=\; z \cdot \frac{s/\sqrt{n}}{\bar{x}}$$

Example B-3
A metering study of 70 CFLs finds the hours of use to average 2.0 per day, with a standard deviation of 0.82 hours. Precision can then be estimated as:

$$\text{Absolute Precision}(\bar{x}) \;=\; 1.645 \cdot \frac{0.82\ \text{hrs/day}}{\sqrt{70}} \;=\; 0.16\ \text{hrs/day}$$

$$\text{Relative Precision}(\bar{x}) \;=\; 1.645 \cdot \left(\frac{0.16\ \text{hrs/day}}{2.15\ \text{hrs/day}}\right) \;=\; 7.5\%$$

Thus, we are 90% confident that average CFL usage is between 1.84 and 2.16 hours per day. Alternately, we can say that the mean hours of use is 2 hours per day, with ±9.8% precision at the 90% confidence level.

[End of Example]

7.3 Using a Ratio Estimator to Estimate a Population Mean

When estimating the population mean of some variable y that is closely correlated with some other variable x—which is known for every member of the population—a ratio estimator should be used to take advantage of the correlation. The known variable x is called an **auxiliary variable**. In energy efficiency evaluations, this is most often seen in realization rates, where the goal is to estimate the *evaluated* savings total, and the program database includes *claimed* savings estimates for each member of the population.

For commercial and industrial projects, *claimed* savings values often incorporate site-specific information, such as square footage of conditioned space and hours of operation. In these cases, *claimed* values vary from project to project and the values can reasonably be expected to correlate with *evaluated* savings values.

The primary interest is in estimating the population mean of some variable y (denoted μ_y), where the variable x_i is known for every member of the population. (Thus, μ_x, the population mean of the x_i, is also known.) Then the ratio-based estimate of μ_y is[27]

$$\hat{\mu}_y = \frac{\sum y_i}{\sum x_i} \cdot \mu_x$$

The ratio estimator is technically biased, but its (unquantifiable) bias will generally be negligible compared to its standard error, provided the sample is not too small (ideally, the sample size should be at least 30). This can be a problem when separate ratio estimators are used for small strata; to avoid this issue savings from small strata should be estimated using a combined stratified ratio estimator, as described in *Appendix C. Sample Design and Weighted Estimates*.

The ratio estimator is similar to the estimator obtained by fitting the regression model $y = bx$. However, software that is not survey-oriented generally does not treat uncertainty correctly for (design-based) ratio estimators.[28] This deficiency is especially pronounced with weighted estimators, because design-based weights describe selection probabilities (see *Appendix C. Sample Design and Weighted Estimates*), whereas ordinary regression weights quantify observation-level standard errors.

The only source of uncertainty in this estimate is the uncertainty in the estimated realization rate,

$$\hat{r} = \frac{\sum y_i}{\sum x_i}$$

Estimator uncertainty is quantified through the standard error. The realization rate's standard error is:[29]

$$\text{Standard error of realization rate} = \widehat{\mathrm{SE}}(\hat{r}) = \frac{1}{\sqrt{n}} \sqrt{\sum \frac{(y_i - \hat{r} \cdot x_i)^2}{\bar{x}^2 \cdot (n-1)}}$$

[27] All summations in this section are taken over the sample, not the population. This point can sometimes lead to confusion when working with ratio estimators.

[28] Sample-based inference, which is based on the selection probabilities inherited from the sample design, is often called *design-based*. By default, regression software usually applies *model-based* inference.

[29] The denominator in this expression uses the sample mean $\bar{x}$, rather than the population mean μ_x. This is consistent with Särndal 1992 (page 181, eq. 5.6.12) and the California Evaluation Protocol, but Lohr 1999 (page 68, eq. 3.7) uses the population mean instead. None of these references explicitly compares the two choices. Both possibilities are mentioned in Cochran 1977 (page 155, eqns. 6.12 and 6.13) and in Thompson 2002 (page 69, eqns. 5 and 7), but neither reference states a clear preference. One reason for our preference is that the standard error could be "gamed" by choosing small-scale projects if the population mean were used.

Thus, the standard error of the ratio-based estimate of μ_y is:

$$\widehat{\mathrm{SE}}(\hat{\mu}_y) \;=\; \widehat{\mathrm{SE}}(\hat{r} \cdot \mu_x) \;=\; \widehat{\mathrm{SE}}(\hat{r}) \cdot \mu_x \;=\; \frac{1}{\sqrt{n}} \cdot \frac{\mu_x}{\bar{x}} \cdot \sqrt{\sum \frac{(y_i - \hat{r} \cdot x_i)^2}{n-1}}$$

To express these standard errors more succinctly, write:

$$s^{(\text{ratio})} = \sqrt{\sum \frac{(y_i - \hat{r} \cdot x_i)^2}{n-1}}$$

Then the expressions become:

$$\widehat{\mathrm{SE}}(\hat{r}) \;=\; \frac{s^{(\text{ratio})}}{\sqrt{n}} \cdot \frac{1}{\bar{x}}$$

$$\widehat{\mathrm{SE}}(\hat{\mu}_y) \;=\; \widehat{\mathrm{SE}}(\hat{r} \cdot \mu_x) \;=\; \frac{s^{(\text{ratio})}}{\sqrt{n}} \cdot \frac{\mu_x}{\bar{x}}$$

To see how ratio-based estimates leverage auxiliary data to increase study efficiency, compare this formula with the standard error of the sample mean in the previous section. The ratio-based standard error only has to account for the portion of variability in the y_i that is not explained by the realization-rate-adjusted x_i.

In cases where the realization rate itself is of primary interest, precision may be best described in absolute terms. However, when a population average (or total) is the estimation target, relative precision is usually needed. Depending on context, the precision is calculated with one of the following expressions.

$$\text{Absolute Precision}(\hat{r}) \;=\; z \cdot \widehat{\mathrm{SE}}(\hat{r})$$

$$\text{Relative Precision}(\hat{r}) \;=\; z \cdot \frac{\widehat{\mathrm{SE}}(\hat{r})}{\hat{r}}$$

$$\text{Relative Precision}(\hat{\mu}_y) \;=\; z \cdot \frac{\widehat{\mathrm{SE}}(\hat{\mu}_y)}{\hat{\mu}_y} \;=\; z \cdot \frac{\widehat{\mathrm{SE}}(\hat{r} \cdot \mu_x)}{\hat{r} \cdot \mu_x} \;=\; z \cdot \frac{\widehat{\mathrm{SE}}(\hat{r})}{\hat{r}}$$

Note that the relative precision of the estimated *evaluated* mean, $\hat{\mu}_y = \hat{r} \cdot \mu_x$, is exactly the same as the relative precision of the realization rate, $\hat{r}$. This is because $\widehat{\mathrm{SE}}(\hat{r} \cdot \mu_x) = \widehat{\mathrm{SE}}(\hat{r}) \cdot \mu_x$, so the *evaluated* total's relative precision expression has cancelling factors of μ_x in its numerator and denominator.

Example B-4

In an impact evaluation for a commercial efficiency program, $n = 20$ projects are randomly selected for on-site verification. For each site, we have both *claimed* and *evaluated* savings estimates.[30] The *claimed* total for the sampled sites is 607,415 kWh and the *evaluated* total for the sampled sites is 745,104 kWh, so the estimated realization rate is 1.227.

The data and the line $y = 1.227 \cdot x$ are plotted in Figure 2.

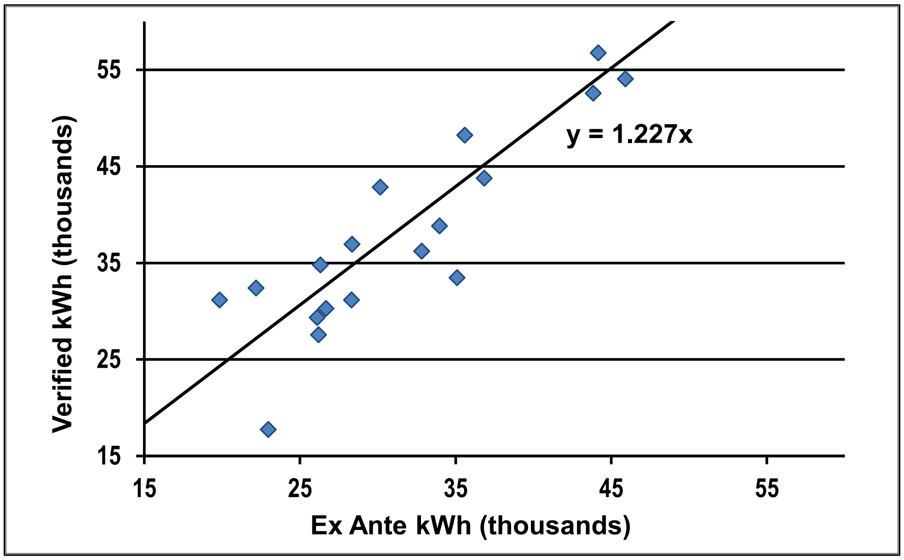


Figure 2. Verified versus claimed savings values

For these data, $s^{(\text{ratio})} = 6{,}176$ kWh and $\bar{y} = 39{,}216$ kWh. Thus, at the 90% confidence level, the relative precision is:

$$\text{Relative Precision}(\hat{r} \cdot \mu_x) \;=\; 1.645 \cdot \frac{6{,}176 \,/\sqrt{20}}{39{,}216} \;=\; 5.8\%$$

If we ignored the auxiliary (*claimed*) data and used the sample mean estimator, $N \cdot \bar{y}$, instead of the ratio estimator, we would need to replace $s^{(\text{ratio})}$ with the standard deviation of the sample's verified savings numbers (in this case, $s = 12{,}132$ kWh). We would then obtain this:

$$\text{Relative Precision}(N \cdot \bar{y}) \;=\; 1.645 \cdot \frac{12{,}132 \,/\sqrt{20}}{39{,}216} \;=\; 11.4\%$$

Here, the ratio estimator's precision is roughly one-half of the mean-based estimator's precision. This is because the ratio estimator's s-factor only needs to account for deviations between verified savings values and realization rate-adjusted *claimed* values $(y_i - \hat{r} \cdot x_i)$. However, the mean-based s-factor (the usual sample standard deviation) must account for deviation between each verified savings value and the mean of the verified savings values $(y_i - \bar{y})$.

Figure 3 shows the spread of the two types of deviations for this example.

[30] Claimed values are the values in the program database, and evaluated values are engineering estimates based on data collected on-site during the evaluation.

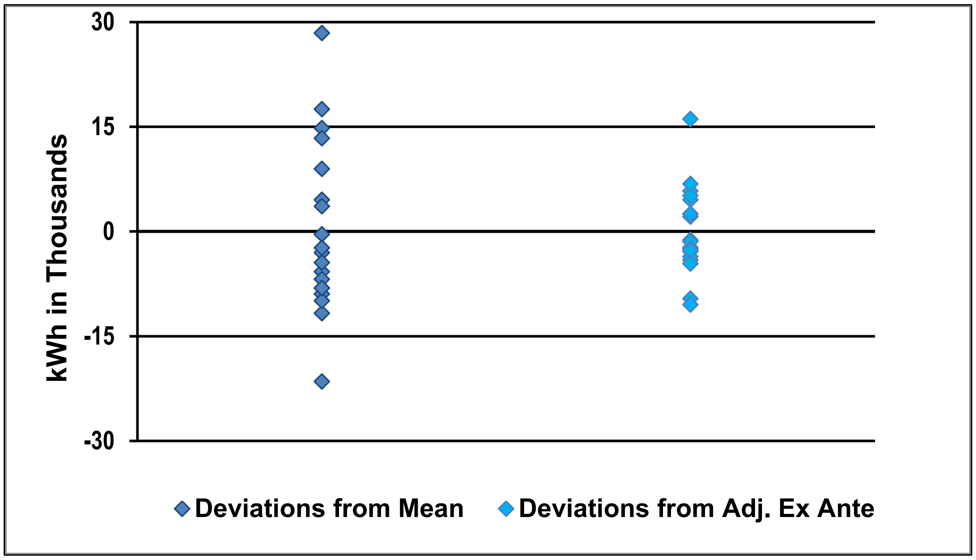


Figure 3: Comparison of verified savings deviations

[End of Example]

To develop intuition, it is helpful to think of the sizes of $s^{(\text{ratio})}$ and s relative to $\bar{y}$, rather than in absolute terms. Example B-4 had $s^{(\text{ratio})}/\ \bar{y} = 15.7\%$ and $s/\ \bar{y} = 30.9\%$. The expression $s^{(\text{ratio})}/\ \bar{y}$ is called the **error ratio** (ER), and $s/\ \bar{y}$ is the **coefficient of variation** (CV). These quantities describe the typical deviation size as a percentage of the typical project size.

In general, the deviations captured by $s^{(\text{ratio})}$ and s may reflect a number of unpredictable factors. For $s^{(\text{ratio})}$, the deviations between verified savings and adjusted *claimed* savings may result from factors such as poor data handling at the time of implementation, changes in site conditions since implementation, or changes in the number of shifts operating at the site. The standard deviation s may be influenced any of these factors, plus general variability among project sizes. As a result, the ER and CV do not obey any firm rules, except that the ER will generally be smaller than the CV whenever verified savings is roughly proportional to *claimed* savings.[31] (Also, most evaluators would agree that an ER of 15.7% and a CV of 30.9% are quite small for a commercial program.)

Example B-5

The program database for a commercial gas efficiency program indicates 9.42 million Mcf [thousand cubic feet] of claimed (*claimed*) savings program-wide, so we will conduct 40 site visits to verify the claimed savings. The 40 sampled sites account for a total of 2.00 mMcf in claimed savings, and our site visits verify a total of 1.70 mMcf in savings. Then we have:

[31] In general, the ratio estimator will be more efficient than the mean-based estimator if the correlation between x and y is greater than $0.5 \cdot \text{CV}(x) \cdot \text{CV}(y)$ (Cochran, 1977, page 157).

$$\begin{array}{lllll} \hat{r} & = & 1.7 \text{ mMcf}/2.0 \text{ Mcf} & = & 85.0\% \\ \bar{y} & = & 1.7 \text{ mMcf}/40 & = & 0.0425 \text{ mMcf} \\ \bar{x} & = & 2.0 \text{ mMcf}/40 & = & 0.0500 \text{ mMcf} \end{array}$$

Our data yields $s^{(\text{ratio})} = 0.0233$ mMcf, so the error ratio is:

$$\text{ER} = 0.0233 \text{ mMcf}/0.0425 \text{ mMcf} = 54.8\%$$

At the 90% confidence level, the realization rate's absolute precision is:

$$\text{Absolute Precision}(\hat{r}) \ = \ 1.645 \cdot \frac{s^{(\text{ratio})}}{\sqrt{n}} \cdot \frac{1}{\bar{x}} \ = \ 1.645 \cdot \frac{0.0233}{\sqrt{40}} \cdot \frac{1}{0.05} \ = \ 0.121$$

In other words, we have 90% confidence that the population realization rate is within 12.1 *percentage points* of 85%.

We estimate the program-wide total savings as $0.85 \cdot 9.42$ mMcf $= 8.01$ mMcf.

To calculate the relative precision of this estimate, we use:[32]

$$\text{Relative Precision}(\hat{r} \cdot \mu_x) \ = 1.645 \cdot \frac{s^{(\text{ratio})}/\sqrt{n}}{\bar{y}} \ = \ 1.645 \cdot \frac{0.0233/\sqrt{40}}{0.0425} \ = \ 14.3\%$$

So, we are 90% confident that the actual program savings is within 14.3% percent of 8.02 mMcf.

If we ignored the auxiliary (*claimed*) data and used the sample mean estimator, $N \cdot \bar{y}$, instead of the ratio estimator, we would have to replace the error ratio, $s^{(\text{ratio})}/\bar{y} = 54.8\%$, with the coefficient of variation, $s/\bar{y}$.

As noted earlier, the CV will be greater than the ER when *evaluated* and *claimed* values are strongly correlated. For example, if the CV in this example is 93.1%, then the mean-based estimator would be much less precise:

$$\text{Relative Precision}(N \cdot \bar{y}) \ = \ 1.645 \cdot 0.931 \cdot \frac{1}{\sqrt{40}} \ = \ 24.2\%$$

[End of Example]

[32] Recall that the relative precision of the population total estimate is the same as the relative precision of the population mean estimate, because both of the estimates and their standard errors differ by a factor of N from one setting to the other.

7.4 Estimating a Difference or Sum

- Sums and differences of estimated quantities arise frequently in evaluation work. Two prominent examples are:
- ***Combining savings across domains or strata.*** Large studies are often composed of multiple distinct research tasks for which the savings from the various research domains are to be summed to estimate the composite savings.
- ***Calculating savings as a difference.*** Savings is the difference between consumption in an inefficient scenario and consumption in an efficient one. Because energy efficiency evaluations seek to estimate these savings, evaluators often need to estimate a difference rather than a mean or proportion.

Assume independent, unbiased estimates, $\hat{x}$ and $\hat{y}$, of target quantities x and y. The difference or sum of the two estimates is an unbiased estimate of the difference or sum of the targets:

$$\widehat{x \pm y} = \hat{x} \pm \hat{y}$$

The standard error of the estimated difference or sum is then a function of both estimators. In general, this is:

$$\mathrm{SE}(\hat{x} \pm \hat{y}) = \sqrt{\mathrm{SE}(\hat{x})^2 + \mathrm{SE}(\hat{y})^2 + 2 \cdot \mathrm{Cov}(\hat{x}, \hat{y})}$$

Here, $\mathrm{Cov}(\hat{x}, \hat{y})$ is the covariance of the two estimators. When the two estimators are based on separate, independently drawn samples, their sampling errors will be independent and their covariance will equal zero. In such cases, the formula reduces to:

$$\mathrm{SE}(\hat{x} \pm \hat{y}) = \sqrt{\mathrm{SE}(\hat{x})^2 + \mathrm{SE}(\hat{y})^2}$$

When the sampling errors are not independent, the evaluator will either need to estimate the covariance[33] or employ an alternate method, such as the bootstrap.

The absolute and relative precision are then estimated as:

$$\text{Absolute Precision}(\hat{x} \pm \hat{y}) = z \cdot \widehat{\mathrm{SE}}(\hat{x} \pm \hat{y})$$

$$\text{Relative Precision}(\hat{x} \pm \hat{y}) = z \cdot \left(\frac{\widehat{\mathrm{SE}}(\hat{x} \pm \hat{y})}{\hat{x} \pm \hat{y}} \right)$$

Example B-6

A utility ran a CFL program and a refrigerator-recycling program, so the evaluator randomly sampled 30 projects from the CFL program and independently sampled 35 projects from the recycling program. The CFL sample led to an estimated program savings of 20 GWh, and the

[33] The procedure for evaluating the covariance will depend on the particular estimators and their relationship to one another.

refrigerator-recycling program had an estimated savings of 5 GWh. The total portfolio savings was then estimated as 25 GWh.

Assume both program-level estimators had 10% relative precision at the 90% confidence level. To evaluate the uncertainty of total savings, we first calculate the standard error for each program:

$$\widehat{\mathrm{SE}}(\text{CFL Savings}) = \frac{10\% \cdot 20\text{ GWh}}{1.645} = 1.22\text{ GWh}$$

$$\widehat{\mathrm{SE}}(\text{Refrigerator Savings}) = \frac{10\% \cdot 5\text{ GWh}}{1.645} = 0.30\text{ GWh}$$

The total program relative precision is then:

$$\text{Relative Precision}(\text{Portfolio Savings}) = \frac{1.645 \cdot \sqrt{(1.22)^2 + (0.30)^2}}{20 + 5} = 8.2\%$$

[End of Example]

7.5 Estimating a Product

In some instances, the product of two estimates is required. A common example of this is in using installation rates, where the proportion of measures installed is multiplied by an estimated per-unit savings to arrive at final verified savings.

In general, the exact standard error of a product is quite complicated,[34] but when the two estimators' sampling errors are independent, the standard error is:

$$\mathrm{SE}(\hat{x} \cdot \hat{y}) = \sqrt{(\hat{x} \cdot \mathrm{SE}(\hat{y}))^2 + (\hat{y} \cdot \mathrm{SE}(\hat{x}))^2 + (\mathrm{SE}(\hat{x}) \cdot \mathrm{SE}(\hat{y}))^2}$$

[34] The delta method yields a reasonably simple approximation that includes a covariance term. However, in evaluation work, there are few circumstances in which a product of two non-independent estimators is needed. In these rare cases, one should either apply the bootstrap method or, if the covariance can be estimated, the delta method.

Example B-7

For an evaluation of an HVAC program, the estimated gross annual unit energy savings is 200 kWh, with a standard error of 12.2 kWh/year. (This corresponds to 10% relative precision.)

The client and regulator have agreed that net savings will be calculated using the net-to-gross (NTG) ratio from a previous year's evaluation. The earlier evaluation reported an NTG estimate of 80% with a SE of 3.2% (absolute precision) at the 90% confidence level. Net unit savings is then estimated as $200\text{ kWh} \cdot 0.8 = 160\text{ kWh}$ per year.

Because the NTG estimate is independent of the gross estimate, the relative precision of net per-unit savings is:

$$\frac{1.645\sqrt{(80\% \cdot 12.2)^2 + (200 \cdot 3.2\%)^2 + (12.2 \cdot 3.2\%)^2}}{160} = 12.0\%$$

Note that the net savings estimate is less precise than the gross savings estimate (12% versus 10% relative precision, respectively). This is due to the additional uncertainty introduced through the NTG factor.

[End of Example]

7.6 Summary of Analytical Techniques

Table 8 summarizes the basic formulas used for analysis of simple random samples.

Table 8. Sample Analysis Formulas for Large Populations

Estimator and Target Quantity	Expression	Standard Error	Data Type
Sample proportion ($\hat{p}$); Population proportion (p)	$\frac{n_x}{n}$	$\frac{1}{\sqrt{n}} \cdot \sqrt{\hat{p}(1-\hat{p})} = \frac{s^{(p)}}{\sqrt{n}}$	Binomial
Sample mean ($\bar{x}$); Population mean (μ_x)	$\frac{\sum x_i}{n}$	$\frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\sum(\bar{x}-x_i)^2}{n-1}} = \frac{s}{\sqrt{n}}$	Quantitative
Ratio estimator ($\hat{r} \cdot \mu_x$); Population mean (μ_y)	$\frac{\sum y_i}{\sum x_i} \cdot \mu_x$	$\frac{1}{\sqrt{n}} \cdot \sqrt{\frac{\sum(y_i - \hat{r}x_i)^2}{n-1}} \cdot \frac{\mu_x}{\bar{x}} = \frac{s^{(\text{ratio})}}{\sqrt{n}} \cdot \frac{\mu_x}{\bar{x}}$	Quantitative
Sum (or difference)*	$\hat{x} \pm \hat{y}$	$\sqrt{\mathrm{SE}(\hat{x})^2 + \mathrm{SE}(\hat{y})^2}$	Either
Product*	$\hat{x} \cdot \hat{y}$	$\sqrt{(\hat{x} \cdot \mathrm{SE}(\hat{y}))^2 + (\hat{y} \cdot \mathrm{SE}(\hat{x}))^2 + (\mathrm{SE}(\hat{x}) \cdot \mathrm{SE}(\hat{y}))^2}$	Either

*The indicated standard error formula is only valid if estimators are statistically independent (see the previous two subsections).

8 Appendix C. Sample Design and Weighted Estimates

For the estimators in Appendix B, it was assumed the sample was drawn through simple random sampling from a large population. This section discusses estimation with more general sample designs. Much of the discussion focuses on stratified designs and related topics, such as weighted estimators and sample optimization. We also discuss sampling with probability proportional to size and two-stage sampling for assessing savings for large projects.

8.1 Simple Random Sampling

In many ways, simple random sampling (SRS) is the most natural and intuitive sample design. In fact, more complicated designs can often be thought of as modifications or combinations of SRS. As the name suggests, SRS without replacement is the simplest random sampling approach, equivalent to "drawing n names from a hat."[35] The defining feature is that the final sample could be any set of n distinct names, and all such sets are equally likely. Thus, for an SRS of size n from a population of size N, each individual unit has selection probability n/N.

8.1.1 Sample Means with FPC

The only difference between this section and the sample mean discussion in Appendix B is that a very large population is no longer assumed.

Example C-1

For estimating the average number of incandescent bulbs still operating in residences within some utility's territory, the estimation target is the population mean,

$$\mu_x = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Here,

N = utility's total number of residential customers (the population size)

x_i = the number of incandescent bulbs operating at the i^{th} residence.

To estimate μ_x, we directly verify the number of incandescent bulbs in each of n homes, where the homes are selected via SRS. Based on these data, the most natural estimate of μ_x is the sample mean:

$$\bar{x} = \frac{1}{n} \sum_{\text{sampled } i} x_i$$

[35] The names are drawn without replacement, which means once a name is drawn, it is excluded from subsequent selection rounds. Thus, no name can be drawn more than once.

The standard error of the sample mean of an SRS is:

$$\widehat{\mathrm{SE}}(\bar{x}) = \sqrt{1-\frac{n}{N}} \cdot \frac{1}{\sqrt{n}} \cdot \sqrt{\sum_{\text{sample}} \frac{(x_i-\bar{x})^2}{(n-1)}} = \sqrt{1-\frac{n}{N}} \cdot \frac{s}{\sqrt{n}}$$

[End of Example]

Readers who are familiar with the statistical properties of sample means but not familiar with finite population inference may be surprised by the factor of $\sqrt{1-n/N}$ in the standard error expression.

This is called the **finite population correction (FPC)**, and it is a direct result of the SRS sample design. The FPC can be thought of as accounting for the fact that when the sample represents a significant fraction of the population, the uncertainty about the population mean is reduced. Note that when the population size is very large compared to the sample size, the ratio n/N will be close to zero, so the FPC will be close to one. In other words, the FPC is negligible for large populations.[36] In contrast, when the sample size is large so that n/N is close to one, the FPC (and hence the standard error) will be close to zero. A very large sample size means that most of the population has been measured directly, leaving little uncertainty about the population mean.

Determining an appropriate sample size is a critical step in planning a study. This determination is generally based on an agreed-upon precision target and some fixed confidence level. The general procedure uses the relevant precision formula and the target precision and confidence levels to express the necessary sample size in terms of important population quantities.

For the sample mean under SRS, the relative precision formula is typically used:

$$\text{Relative Precision}(\bar{x}) = \mathrm{z} \cdot \frac{\widehat{\mathrm{SE}}(\bar{x})}{\bar{x}}$$

The simplest way to calculate the sample size proceeds in two steps:

1. Calculate an initial sample size, $\boldsymbol{n_0}$, using the large-population standard error formula (that is, the formula without the FPC).
2. Adjust the initial sample size to account for the FPC in the true standard error.

The next example illustrates Step 1 and is followed by a brief discussion of the parameters that drive sample sizes. Step 2 is discussed at the end of this section.

Example C-2

To estimate the population mean to within 10% of its true value with 90% confidence, Step 1 ignores the FPC to obtain the initial sample size, n_0. This is the smallest integer that yields

[36] The proportion, sample mean, and ratio estimator sections of *Appendix B* provided standard error formulas that are valid under the assumption that the FPC is negligible.

$$0.10 \geq 1.645 \cdot \frac{s/\sqrt{n_0}}{\bar{x}}$$

Equivalently, n_0 is the smallest integer that satisfies this equation:

$$n_0 \geq \left(\frac{1.645}{0.10}\right)^2 \cdot \left(\frac{s}{\bar{x}}\right)^2$$

The quantity $s/\bar{x}$ is called the sample **coefficient of variation** (CV). This factor will not be known until after the data are collected. Past experience is the best guide for determining plausible values for the CV.

If the sample-based CV is greater than was expected when the sampling plan was developed, the study will fail to meet the agreed-upon confidence/precision target. For large studies, it may be advisable to (1) conduct a pilot study to estimate the CV in advance of the primary data collection effort or (2) plan for staged data collection so that sample sizes for later stages can be adjusted to reflect the CV observed through earlier stages. In all cases, the evaluator and the client should agree in advance on the measures to be taken to ensure an adequate sample size.

[End of Example]

As shown in the calculation in Example C-2, the large-population sample size formula is:

$$n_0 = \left(\frac{z \cdot \mathrm{CV}}{e_{\mathrm{rel.}}}\right)^2$$

Where:

CV is the coefficient of variation, the standard deviation divided by the mean
$e_{\mathrm{rel.}}$ is the desired level of relative precision
z is the critical value of the standard normal distribution value for the desired confidence level

For example, for 90% confidence, 10% precision, and a CV of 0.5, the initial sample size is:

$$n_0 = \left(\frac{1.645 \cdot 0.5}{0.10}\right)^2 = 67.7$$

Therefore, a sample of size 68 should be used here if the FPC is negligible. (Researchers often assume a CV of 0.5 when determining sample sizes, and because 90/10 confidence/precision is a common target, samples of size 68 are very common.)

One reason CVs of 0.5 are often reasonable in evaluation work is that the savings values are typically positive for all (or nearly all) projects. If 95% of a program's projects have savings between zero and 200% of the mean savings, and if the savings values are approximately normally-distributed, then a CV of 0.5 will apply.[37] This value, however, should not be applied

[37] Recall that for a normal distribution, approximately 95% of the population will fall within two standard deviations (SD) of the mean. If the CV equals 0.5, then the SD is one half of the mean. Thus, the 95% interval, mean ± 2

without due consideration of the expected nature of program savings. The justification noted here does not apply if project savings are heavily skewed towards large savers (in this case, the normality assumption fails). A stratified design (described later in this appendix) can often resolve this sort of skew and yield an effective CV that is closer to 0.5. In general, comparable previous studies and evaluation experience are the best guides for assessing likely CV values.

Because the FPC reduces standard error, it also reduces sample size required for any fixed levels of precision and confidence and fixed CV. The finite population adjustment reduces the necessary sample size as follows:

$$n = \frac{n_0 \cdot N}{n_0 + N}$$

In Example C-2, if the target population is of size $N = 200$, then the population is only three times the size of the sample. In this case, the finite population adjustment reduces the required sample size from 68 to 50:

$$n = \frac{68 \cdot 200}{68 + 200} \approx 50$$

8.1.2 Population Proportions and Ratio Estimators With FPC

Proportion estimates and ratio estimates can both be interpreted as versions of sample means. Thus, under SRS, these estimators' standard errors and sample sizes undergo finite population adjustments that are identical to their sample mean analogues.

The estimators themselves are unchanged from the large population case:

$$\begin{aligned} \hat{p} &= \frac{n_x}{n} \\ \hat{r} \cdot \mu_x &= \frac{\sum y_i}{\sum x_i} \cdot \mu_x \end{aligned}$$

Their standard errors, however, are multiplied by a finite population correction, just as in the sample mean case:

$$\begin{aligned} \widehat{\mathrm{SE}}(\hat{p}) &= \sqrt{1 - \frac{n}{N}} \cdot \frac{\sqrt{\hat{p} \cdot (1 - \hat{p})}}{\sqrt{n}} \\ \widehat{\mathrm{SE}}(\hat{r} \cdot \mu_x) &= \sqrt{1 - \frac{n}{N}} \cdot \frac{1}{\sqrt{n}} \cdot \sqrt{\sum \frac{(y_i - \hat{r} \cdot x_i)^2}{n - 1}} = \sqrt{1 - \frac{n}{N}} \cdot \frac{s^{(\text{ratio})}}{\sqrt{n}} \cdot \frac{\mu_x}{\bar{x}} \end{aligned}$$

SD, is the same as mean ± mean (the mean, plus or minus itself). In other words, if the CV is 0.5 and the data are normal, the 95% CI will range from 0 to 200% of the mean. Again, if one is willing to assert that the data will be normal and that most of the members of the population will fall between 0 and 200% of the mean, then a CV of 0.5 is appropriate.

Sample size calculations for both population proportions and ratio estimators are similar to the sample mean calculations. Calculate an initial sample size, n_0, using the large-population standard error formula and then apply a finite population adjustment.

For population proportions the large-population precision formula is:

$$e_{\text{abs.}} = z \cdot \widehat{\text{SE}}(\hat{p}) = z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n_0}}$$

So the initial sample size formula is:

$$n_0 = \left(\frac{z}{e_{\text{abs.}}}\right)^2 \cdot p(1-p)$$

In this formula, z is as before and $e_{\text{abs.}}$ is the absolute precision target. If there is no basis for making *a priori* assumptions about p, then use $p = 0.5$, because $p(1-p)$ obtains its maximum with this value.

For both population proportions and ratio estimators, the FPC reduces the necessary sample size as before. In both cases, the final sample size is:

$$n = \frac{n_0 \cdot N}{n_0 + N}$$

Example C-3

For a large population, the requirement for estimating a population proportion to within 5 percentage points, with 90% confidence, is this:

$$n_0 \geq \left(\frac{1.645}{0.05}\right)^2 \cdot p(1-p)$$

The quantity $p(1-p)$ can never be greater than $0.5(1-0.5) = 0.25$, so the precision target is guaranteed to be met if:

$$n_0 \geq \left(\frac{1.645}{0.05}\right)^2 \cdot (0.5)^2 = 270.6$$

Thus, if the population is very large and there is no *a priori* knowledge of *p*, then to meet the 90/5 standard, plan for the study to achieve at least 271 complete responses.

Now assume there are only $N = 550$ individuals in the target population. Then the FPC reduces the required sample size to:

$$n = \frac{270.6 \cdot 550}{270.6 + 550} = 181.4$$

In this case, plan for 182 complete survey responses.

[End of Example]

When the ratio estimator $\hat{r} \cdot \mu_x$ is used to estimate the population mean μ_y, the large-population precision formula is:

$$e_{\text{rel.}} \;=\; z \cdot \frac{\widehat{\text{SE}}(\hat{r} \cdot \mu_x)}{\hat{r} \cdot \mu_x} \;=\; z \cdot \frac{s^{(\text{ratio})}/\sqrt{n_0}}{\bar{y}}$$

Therefore, the initial sample size formula is:

$$n_0(\hat{r} \cdot \mu_x) \;=\; \left(\frac{z}{e_{\text{rel.}}}\right)^2 \left(\frac{s^{(\text{ratio})}}{\bar{y}}\right)^2$$

This formula is identical to the one obtained for the sample mean, except that the standard deviation, s, has been replaced with $s^{(\text{ratio})}$, which quantifies only that portion of variability not explained through the auxiliary information.

The quantity $s^{(\text{ratio})}/\bar{y}$ is called the **error ratio (ER)**.[38] When the x and y variables are correlated, the error ratio will tend to be smaller than the CV, so the ratio-based estimator will be more efficient than the sample mean.

As indicated above, the FPC reduces the necessary sample size precisely as before. In both cases, the final sample size is:

$$n = \frac{n_0 \cdot N}{n_0 + N}$$

[38] The California Evaluation Framework prescribes a model-assisted approach, based on evidence that deviations between evaluated values y_i and adjusted claimed values $\hat{r}x_i$ tend to scale in proportion to x_i^{γ} for some $\gamma \approx 0.8$. This approach leads to a different procedure for estimating the error ratio. When greater efficiency may be gained through this well-studied model-based approach, researchers are encouraged to apply it.

Summary of SRS Estimators

The important equations for SRS are listed in Table 9.

Table 9. Results for Simple Random Samples

Estimator	Expression	Standard Error	Initial Sample Size	Sample Size With FPC
Sample mean	$\frac{\sum x_i}{n}$	$\sqrt{1-\frac{n}{N}}\cdot\frac{s}{\sqrt{n}}$	$n_0 = \left(\frac{z}{e_{\text{rel.}}}\right)^2 \cdot (\text{CV})^2$	$\frac{n_0 \cdot N}{n_0 + N}$
Sample proportion	$\frac{n_x}{n}$	$\sqrt{1-\frac{n}{N}}\cdot\frac{\sqrt{p(1-p)}}{\sqrt{n}}$	$n_0 = \left(\frac{z}{e_{\text{abs.}}}\right)^2 \cdot p(1-p)$	$\frac{n_0 \cdot N}{n_0 + N}$
Ratio estimator	$\frac{\sum y_i}{\sum x_i}\cdot \mu_x$	$\sqrt{1-\frac{n}{N}}\cdot\frac{s^{(\text{ratio})}}{\sqrt{n}}\cdot\frac{\mu_x}{\bar{x}}$	$n_0 = \left(\frac{z}{e_{\text{rel.}}}\right)^2 \cdot (\text{ER})^2$	$\frac{n_0 \cdot N}{n_0 + N}$

8.2 Stratified Random Sampling

Stratified sampling entails partitioning the population into distinct groups (called *strata*) and drawing samples independently from each stratum. In some cases, the groupings reflect qualitative population characteristics. For example, participants in a commercial HVAC program may be stratified by business type, or participants in a comprehensive nonresidential program may be separated by custom versus prescriptive projects. Strata may also be created to group the population into size categories according to *claimed* savings values in the program database.

The main reason for using stratified sampling is to reduce the variance in a population-wide estimator by separating the population into homogeneous groups. Population-level uncertainty is then driven exclusively by within-stratum variation. As a result, when homogeneous groupings are available, stratified random sampling is almost always more efficient than simple random sampling. In addition, in cases of study domains with particularly small populations, stratification ensures that every relevant stratum is represented in the sample. (This may not be case in simple random sampling.)

Stratification is a very flexible tool in its application. For instance, the population of program participants may first be divided into sector and fuel type groupings and then stratified by size. The particular choice of stratification variable(s) will depend on context.

For this section, assume that (1) the population has been partitioned into *H* non-overlapping strata and (2) the stratum population sizes are given by $N_1, N_2, \dots, N_H$. Also assume that each stratum's sample is selected via simple random sampling within the stratum.[39] For example, within stratum h, an SRS of size n_h is been drawn from a group of N_h individuals, so each

[39] Stratification can also be employed with more general probability sampling within each stratum. (This is described in most sample design textbooks.) When an alternative scheme is used, the researcher should clearly describe the sampling scheme and the estimator with references (or direct calculations) explaining why standard error calculations are valid indicators of uncertainty.

sampled unit represents N_h/n_h members of the population. Thus, the weight of a unit sampled from stratum h is $w_h = N_h/n_h$.

Stratified designs bring new notational requirements. For most objects, a subscripted h will indicate stratum number, and a subscripted *all* will indicate that an object spans all strata. Most stratified approaches are more easily understood when research tasks are expressed in terms of population totals (and their estimators) rather than population means, so the notation also makes this distinction.

The general conventions for this section are as follows.

Population Quantities

X_{all} and Y_{all} are the x_i and y_i population totals
N_{all} is the total number of population members, $N_{\text{all}} = N_1 + N_2 + \cdots + N_H$
μ_{all} is the population mean of the x_i, $\mu_{\text{all}} = X_{\text{all}}/N_{\text{all}}$
X_h and Y_h are stratum-h population totals of the x_i and y_i
$\mu_{x,h}$ is the stratum-h population mean of the x_i, $\mu_{x,h} = X_h/N_h$

Sample Quantities and Estimators

n_{all} is the total sample size, $n_{\text{all}} = n_1 + n_2 + \cdots + n_H$
$\bar{x}_h$ and $\bar{y}_h$ are the stratum-h sample means of the x_i and y_i
$w_h = N_h/n_h$ is the weight that applies to stratum-h sample members
$\bar{x}_{\text{all}}^{(\text{w})}$ and $\bar{y}_{\text{all}}^{(\text{w})}$ are the weighted sample means of the x_i and y_i
$h(i)$ is the stratum containing unit i

As before, the procedures for determining appropriate sample sizes will be demonstrated after the basic properties of the estimators are established. Stratified versions of sample means, proportions, and ratio estimators are described in this section.

8.2.1 Stratified Means

The basic idea behind the independent-estimators approach is illustrated in the following example.

Example C-4

For this evaluation, the object is to estimate the total air-conditioning tonnage among all commercial retailers in a particular service territory. A sample mean applied to a simple random sample would be very inefficient, because a small number of commercial retailers are orders of magnitude larger than most of the population. (This skew would translate to a very large CV.)

If retailer size categories are known through auxiliary data, these size categories may be used as strata for the study. Within each stratum, skew would be limited, so stratum-level CVs should be moderate.

Assume three retailer size categories: stratum one covers small retailers, stratum two covers medium retailers, and stratum three covers large retailers. Write s_1 for the stratum-one sample standard deviation, and likewise for $s_2, \ldots, s_H$. Then the estimated stratum one total is $\hat{X}_1 = N_1 \cdot$

$\bar{x}_1$, and its standard error is:

$$\mathrm{SE}(\hat{X}_1) = \mathrm{SE}(N_1 \cdot \bar{x}_1) \ = N_1 \cdot \sqrt{1 - \frac{n_1}{N_1}} \cdot \frac{s_1}{\sqrt{n_1}}$$

Calculate $\hat{X}_2$ and $\hat{X}_3$ the same way, and estimate the population total as:

$$\hat{X}_{\text{all}}^{(\text{w})} = \ \hat{X}_1 + \hat{X}_2 + \hat{X}_3 \ = \ N_1 \cdot \bar{x}_1 + \ N_2 \cdot \bar{x}_2 + N_3 \cdot \bar{x}_3$$

The superscripted “w” emphasizes that this is a weighted estimator. Its standard error is:

$$\mathrm{SE}\left(\hat{X}_{\text{all}}^{(\text{w})}\right) \ = \ \sqrt{\mathrm{SE}(\hat{X}_1)^2 + \mathrm{SE}(\hat{X}_2)^2 + \mathrm{SE}(\hat{X}_3)^2}$$

To estimate the population-wide mean, use:

$$\hat{X}_{\text{all}}^{(\text{w})}/(N_1 + \ N_2 + N_3).$$

This estimate’s standard error is:

$$\mathrm{SE}(\hat{X}_{\text{all}}^{(\text{w})})/(N_1 + \ N_2 + N_3).$$

[End of Example]

The general formula for the stratified-means estimator of the population total is:

$$\hat{X}_{\text{all}}^{(\text{w})} \ = \sum_{h=1}^{H} \hat{X}_h = \sum_{h=1}^{H} N_h \cdot \bar{x}_h$$

This estimator can also be written as a weighted sum,

$$\hat{X}_{\text{all}}^{(\text{w})} \ = \ \sum_{\text{sampled } i} \frac{N_{h(i)}}{n_{h(i)}} \cdot x_i \ = \ \sum_{\text{sampled } i} w_{h(i)} \cdot x_i$$

The weighted sum’s standard error is calculated as follows. (Notice that only the within-stratum standard deviations, s_h, affect the standard error.)

$$\mathrm{SE}(\hat{X}_{\text{all}}^{(\text{w})}) \ = \sqrt{\sum \mathrm{SE}(\hat{X}_h)^2} \ = \ \sqrt{\sum N_h^2 \cdot \mathrm{SE}(\bar{x}_h)^2} \ = \sqrt{\sum \frac{N_h^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot s_h^2}$$

To estimate the population *mean*, divide the estimated total by the population size:

$$\bar{x}_{\text{all}}^{(\text{w})} = \frac{\hat{X}_{\text{all}}^{(\text{w})}}{N_{\text{all}}}$$

This estimator is called the weighted mean.

8.3 Stratified Proportions

The reasoning in the previous section also applies to population proportions. To estimate the fraction of the population having some particular characteristic, first estimate the total number of individuals with the characteristic and then divide by the population size.

To express these results, we must expand on the notation of Appendix B:

> N_{all}^x is the total number of individuals in the population who have characteristic *x*.
> N_h^x is the total number of individuals from stratum h who have characteristic *x*.
> p_{all} is the population proportion, $p_{\text{all}} = N_{\text{all}}^x/(N_1^x + N_2^x + \cdots + N_H^x)$
> n_h^x is the number of *sampled* individuals from stratum h who have characteristic *x*.
> $\hat{p}_h$ is the proportion of the stratum h sample with the characteristic, $\hat{p}_h = n_h^x/n_h$.
> $\hat{p}_{\text{all}}^{(\text{w})}$ and $\widehat{N}_{\text{all}}^x$ are our estimates of p_{all} and N_{all}^x.

The weighted estimators related to population proportions are:

$$\widehat{N}_{\text{all}}^x = \sum_{h=1}^{H} N_h \cdot \hat{p}_h$$

$$\widehat{\text{SE}}\left(\widehat{N}_{\text{all}}^x\right) = \sqrt{\sum N_h^2 \cdot \widehat{\text{SE}}(\hat{p}_h)^2} = \sqrt{\sum \frac{N_h^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \hat{p}_h(1 - \hat{p}_h)}$$

$$\hat{p}_{\text{all}}^{(\text{w})} = \frac{\widehat{N}_{\text{all}}^x}{N_1 + N_2 + \cdots + N_H} = \frac{\sum N_h \cdot \hat{p}_h}{N_1 + N_2 + \cdots + N_H}$$

$$\widehat{\text{SE}}\left(\hat{p}_{\text{all}}^{(\text{w})}\right) = \frac{\widehat{\text{SE}}\left(\widehat{N}_{\text{all}}^x\right)}{N_1 + N_2 + \cdots + N_H}$$

8.3.1 Stratified Ratio Estimators

The stratified ratio estimator is based on the ratio of the weighted sum of the sampled y_i to the weighted sum of the sampled x_i. Rather than applying a different realization rate within each stratum, we apply this single weighted realization rate to all strata. In the preceding section on stratified means, $\hat{X}_{\text{all}}$ represented the weighted total of the x_i, and the weighted mean was $\bar{x}_{\text{all}}^{(\text{w})} = \hat{X}_{\text{all}}/N_{\text{all}}$.

The weighted realization rate can be thought of either as the ratio of estimated totals or as the ratio of estimated means:

$$\hat{r}_{\text{all}}^{(\text{w})} = \frac{\sum_{\text{sample}} w_{h(i)} \cdot y_i}{\sum_{\text{sample}} w_{h(i)} \cdot x_i} = \frac{\hat{Y}_{\text{all}}^{(\text{w})}}{\hat{X}_{\text{all}}^{(\text{w})}} = \frac{\bar{y}_{\text{all}}^{(\text{w})}}{\bar{x}_{\text{all}}^{(\text{w})}}$$

The ratio-based estimate of the population total of the y_i is:

$$\hat{Y}_{\text{all}}^{(\text{w})} = \hat{r}_{\text{all}}^{(\text{w})} \cdot X_{\text{all}} = \frac{\bar{y}_{\text{all}}^{(\text{w})}}{\bar{x}_{\text{all}}^{(\text{w})}} \cdot X_{\text{all}}$$

The standard error is:[40]

$$\begin{aligned}
\text{SE}\left(\hat{Y}_{\text{all}}^{(\text{w})}\right) &= \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right) \cdot \sqrt{\sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \sum_{\substack{\text{stratum } h \\ \text{sample}}} \frac{\left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2}{n_h - 1}} \\
&\approx \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right) \cdot \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{n_h}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \sum_{\substack{\text{stratum } h \\ \text{sample}}} \left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2} \\
&= \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right) \cdot \sqrt{\sum_{\text{sample}} w_{h(i)}\left(w_{h(i)} - 1\right)\left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2}
\end{aligned}$$

Typically, $\mu_{\text{all}}/\bar{x}_{\text{all}}^{(\text{w})}$ will be close to one, because it is the ratio of the actual mean to the estimated mean. So to see the basic features of the standard error formula, we can ignore this factor. What remains in the first equation in the chain above is very similar to the standard error of the weighted sum, $\hat{X}_{\text{all}}^{(\text{w})}$. The only difference is that the s_h^2 of the weighted sum's standard error is now replaced by:

$$\left(s_h^{(r,\text{w})}\right)^2 = \sum_{\substack{\text{stratum } h \\ \text{sample}}} \frac{\left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2}{n_h - 1}$$

The last formula in the chain is identical to the formula provided in the *California Evaluation Framework.* Although the FPC is obscured in the *Framework*'s weight-based presentation, the middle expression clearly shows that the formulation does account for the FPC.

8.3.2 Summary of Estimators for Stratified Samples

The next two tables summarize results for the estimators developed in this section. Table 10 gives the estimators themselves and their standard errors.

[40] See Särndal 1992, page 181.

Table 10. Formulas for Stratified Estimators

Estimator	Expression	Standard Error
Weighted sum	$\hat{X}_{\text{all}}^{(w)} = \sum N_h \cdot \bar{x}_h = N \cdot \bar{x}_{\text{all}}^{(\text{w})}$	$\sqrt{\sum \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) s_h^2}$
Weighted proportion	$\hat{p}_{\text{all}}^{(w)} = \frac{\sum N_h \cdot \hat{p}_h}{\sum N_h}$	$\sqrt{\sum \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \hat{p}_h (1 - \hat{p}_h)}$
Weighted Ratio Estimator	$\hat{Y}_{\text{all}}^{(r,\,\text{w})} = \hat{r}_{\text{all}}^{(r,\,\text{w})} \cdot X_{\text{all}}$	$\sqrt{\sum \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \left(s_h^{(r,\,\text{w})}\right)^2} \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right)$

Table 11 provides supplementary formulas.

Table 11. Additional Formulas

Estimator	Unit-level Standard Deviation Estimates	Other Expressions
Weighted sum	$s_h^2 = \sum_{\text{sample } h} \frac{(x_i - \bar{x}_h)^2}{(n_h - 1)}$	NA
Weighted proportion	$s_{p,h}^2 = \frac{n_h^x}{n_h} \cdot \left(1 - \frac{n_h^x}{n_h}\right)$	$\hat{p}_h = \frac{n_h^x}{n_h}$
Weighted Ratio Estimator	$\left(s_h^{(r,\,\text{w})}\right)^2 = \sum_{\text{sample } h} \frac{\left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2}{n_h - 1}$	$\hat{r}_{\text{all}}^{(\text{w})} = \frac{\bar{y}_{\text{all}}^{(\text{w})}}{\bar{x}_{\text{all}}^{(\text{w})}}$

8.4 Planning and Optimizing Stratified Designs

The basic result in the optimization of stratified designs is called the **Neyman allocation**. Among all possible allocations of the n sample units to the H strata, the lowest overall variance will be achieved if:

$$n_h = n \cdot \left(\frac{N_h \cdot s_h}{N_1 \cdot s_1 + \cdots + N_H \cdot s_H}\right)$$

This formula has one major shortcoming that may render it unacceptable for planning large scale studies—it does not consider cost-efficiency. If units from Stratum 1 are much more expensive to survey than units from Stratum 2, then the cost-optimal sample design should allocate fewer units to the more expensive stratum.

The **cost-weighted Neyman allocation** addresses this concern. Use c_h for the marginal cost of sampling a single unit from stratum *h.* Assume a fixed budget for data collection. Then among all possible resource allocations, the lowest overall variance will be achieved if, for some *n*,

$$n_h = n \cdot \left(\frac{N_h s_h \ / \sqrt{c_h}}{N_1 s_1 \ / \sqrt{c_h} + \cdots + N_H s_H \ / \sqrt{c_h}} \right)$$

Both the Neyman allocation and the cost-weighted Neyman allocation work the same with other estimators. Simply replace the stratum-level standard deviation s_h with the appropriate selection from Table 11.

Table 12. Sample Allocation Formulas

Step	Formula
Estimate maximum acceptable overall variance	$\mathrm{Var}(\hat{X}_{\mathrm{all}}) \ = \ (X_{\mathrm{all}})^2 \cdot \left(\frac{e_{\mathrm{rel.}}}{z}\right)^2$
Allocate sample among strata.	$n_h \ = \ n \cdot \left(\frac{N_h s_h \ / \sqrt{c_h}}{N_1 s_1 \ / \sqrt{c_h} + \cdots + N_H s_H \ / \sqrt{c_h}} \right)$

At the planning stage, of course, data-driven estimates of stratum-level standard deviations are not available. Planning estimates may come from other studies, general past experience, or agreed-upon values based on known database quality standards.[41]

8.5 General Probability Samples and PPS

In simple random sampling without replacement, it was demonstrated that with a sample of size *n* from a population of size *N*, each individual unit has selection probability of:

$$\pi_i = \frac{n}{N}$$

More general sample designs are available, however, such as **probably proportional to size** (PPS). The idea behind PPS is to sample n units from the population, each with probability proportional to its size. Because such a scheme necessarily requires auxiliary information for determining the π_i, the typical auxiliary information notation is used for this section.

> x_i is the auxiliary information for site *i.* (In evaluation work, this is usually the claimed savings estimate from the program database.)
>
> y_i is the variable of primary interest for site *i.*

[41] This is especially relevant for ratio estimators, because large deviations between evaluated and claimed values often reflect problems in the program database, rather than variation in consumer behavior.

The goal is to estimate the population total, $Y = y_1 + \cdots + y_N$.

In practice, auxiliary data (the x_i) are used as a proxy for the true savings sizes (the y_i) in calculating the π_i. Insofar as the x_i are consistently proportional to the y_i, PPS estimation will result in very low standard errors.[42]

Strict PPS can be difficult to implement in a manner that both (1) yields no repeat entries in the sample and (2) produces a sample of fixed size, n.[43] However, there are several available variants that are easy to implement, but loosen one or both of the requirements noted.

The variant called Poisson sampling (illustrated in Example C-5) produces samples with no repeat entries, but with variable sample sizes. This variant does not require size stratification, because project sizes are appropriately accounted for through probability weighting.

Example C-5

Determine the sample size target, *n,* and use the auxiliary data to set selection probabilities.

$$\pi_i \;=\; n \cdot \frac{x_i}{x_1 + x_2 + \cdots + x_N}$$

In a spreadsheet, generate a random number (distributed uniformly between 0 and 1) for each project and then designate each project as sampled if its random number is less than its π_i value.

Then standard estimator of the population total is:

$$\hat{Y} \;=\; \sum_{\text{sampled } i} \frac{y_i}{\pi_i}$$

This estimator's standard error is estimated as:

$$\widehat{\mathrm{SE}}\left(\hat{Y}\right) \;=\; \sqrt{\sum_{\text{sampled } i} (1 - \pi_i)\left(\frac{y_i}{\pi_i}\right)^2}$$

[End of Example]

Other PPS variants are available (see Särndal, *et al.*, pp. 85-99).

[42] The same statement holds for ratio estimators, so PPS does not have any general efficiency advantage over ratio methods. It is only an alternative approach that avoids the need for size stratification and, thus, may be simpler to employ in some contexts (especially for within-site subsampling, which is described in the next section).

[43] See Särndal, *et al.*, pp. 90-7. A principle difficulty is that the second-order inclusion probabilities can be difficult to evaluate for any given scheme that produces the desired first-order probabilities. Advanced statistical software packages (such as STATA and SAS) can draw samples and analyze data for most PPS variants, so these difficulties are not fatal. However, as the algorithms would not be easy to implement in a spreadsheet, these methods may not be practical for field work.

8.6 Two-Stage Sampling for Large Projects

Nonresidential programs often include a small number of very large projects. In many cases, direct evaluation of every measure within a large project would impose an unacceptable burden on the customer. As a result, evaluators must rely on a subsample of measures within each large project in the set of sampled projects. This is called two-stage sampling.[44]

The principles described in the preceding sections apply both to the overall sample and to each subsample. This section explains how to integrate subsample results with the broader program evaluation. Our guidance is similar to that given in ASHRAE Guideline 14.

Example C-6

An industrial energy efficiency program is being evaluated using a stratified design that includes a single stratum for very large projects (designated as stratum *H*). For this example, assume the following: (1) a weighted-sum estimator will be used to combine stratum-level results and (2) all measures at any sampled site that is not a member of the large projects stratum will be directly evaluated.

For each stratum other than stratum *H,* the estimated total savings is:

$$\hat{X}_h = N_h \cdot \bar{x}_h \quad \text{and} \quad \mathrm{SE}(\hat{X}_h) = \sqrt{\frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) s_h^2}$$

For a sampled site *i* within stratum *H*, we do not directly evaluate the savings x_i. Instead, we estimate x_i using verified values $x_{i,1}, x_{i,2}, \dots, x_{i,m}$ for some sample of measures within site *i.* The particular method for estimating x_i based on the sampled $x_{i,j}$ depends on the site-level sample design and evaluation plan. However, in all cases it is possible to calculate the estimate, $\hat{x}_i$, and its standard error, $\mathrm{SE}(\hat{x}_i)$. The total savings estimate for stratum *H* is then:

$$\hat{X}_H \;=\; N_H \cdot \frac{\hat{x}_1 + \hat{x}_2 + \cdots + \hat{x}_{n_H}}{n_H} \;=\; N_H \cdot \bar{\hat{x}}_H$$

The standard error of this estimate includes both the usual sampling error (as with the other $\hat{X}_h$) and within-site sampling errors:

$$\mathrm{SE}(\hat{X}_H) = \sqrt{\frac{N_H^2}{n_H}\left(1 - \frac{n_H}{N_H}\right) s_H^2 \;+ \sum_{\text{sample } H} \mathrm{SE}(\hat{x}_i)^2}$$

[44] The distinguishing feature of two-stage sampling is that a sample of secondary units (for example, measures) is selected within each sampled primary unit (for example, project). *One-stage sampling* refers to the case where all secondary units are selected from each sampled primary unit. *Cluster sampling* is usually synonymous with two-stage sampling, but some textbooks reserve this term for one-stage sampling.
Also, two-*stage* sampling is not the same as two-*phase* sampling, in which a large initial sample is observed through low-cost interactions (for example, phone surveys), and the initial sample data are used to increase efficiency for a small sample involving more expensive interactions (for example, site visits). (Two-phase sampling is discussed in Section 8.7, *Two-Phase [Nested] Sampling*.)

It is not uncommon to conduct a full census of very large sites. In such cases, $n_H = N_H$, so the first term in the standard error is zero. Therefore, the terms $\mathrm{SE}(\hat{x}_i)^2$ are the sole contributors to the estimator's standard error for any census stratum.

As always, the total program savings is estimated as:

$$\hat{X}_{\text{all}}^{(\text{w})} = \sum_{h=1}^{H} \hat{X}_h \quad \text{and} \quad \mathrm{SE}\left(\hat{X}_{\text{all}}^{(\text{w})}\right) = \sqrt{\sum \mathrm{SE}\left(\hat{X}_h\right)^2}$$

[End of Example]

Example C-6 illustrates an important feature of two-stage sampling—each finite population correction applies only to the level at which the relevant sampling occurs. Thus, the FPC due to first-stage sampling applies to program-level estimates, while within-site sampling may lead to FPCs which apply within the $\mathrm{SE}(\hat{x}_i)$.

ASHRAE Guideline 14 presents this same approach, but with a slightly different perspective on the origin of random deviations between the $\hat{x}_i$ and x_i. In Guideline 14, the standard errors of the $\hat{x}_i$ are assumed to account for measurement, modeling, and similar sources of random error.

This section's guidance is compatible with Guideline 14. In general, dominant error sources should always be accounted in the $\mathrm{SE}(\hat{x}_i)$, and the dominant errors may be due to modeling error in one context and sampling error in another, depending on site-level evaluation strategies.

The following example illustrates an important point regarding the proper handling of auxiliary data when site-level sub-sampling is used.

Example C-7

For an industrial energy efficiency program, the evaluator is using a stratified design and has created a single stratum containing the program's largest projects (designated as stratum *H*). The evaluator plans to evaluate savings directly for every measure at sampled sites that are not members of stratum *H*. For this example, assume the evaluator plans to use a weighted ratio estimator to estimate the total program savings.

For a sampled site *i* in stratum *H*, the evaluator uses whatever means are available to estimate y_i efficiently—that is, to minimize $\mathrm{SE}(\hat{y}_i)$.[45] For some sites, this may include within-site ratio estimation or a PPS estimator. In such cases, the evaluator may review *claimed* savings assumptions on site and adjust *claimed* values to reflect actual hours of use and similar inputs, provided that the adjustments are (1) applied to sampled and non-sampled measures alike and (2) based on information that is equally available for sampled and non-sampled measures.

For example, if the *claimed* values in the program database assume a 16-hour daily schedule for every measure at a given site, but the site actually operates for 24 hour per day, the measure-

[45] Recall that for ratio estimators, y_i represents verified savings and x_i represents claimed savings estimates.

level *claimed* values may be adjusted accordingly. The main requirement is that such adjustments be made without giving the site's sampled measures any special consideration.[46]

Also, because *claimed* values cannot be adjusted for every site in the population, this sort of *a priori* adjustment applies only to measures within a sampled site and only to the calculation of $\hat{y}_i$ and $\mathrm{SE}(\hat{y}_i)$. The original *claimed* values must still be used in calculating the program-level standard error.

In this case, the estimated the realization rate is determined as:

$$\hat{r}_{\text{all}}^{(\text{w})} = \frac{N_1 \cdot \bar{y}_1 + N_2 \cdot \bar{y}_2 + \cdots + N_{H-1} \cdot \bar{y}_{\text{H-1}} + N_H \cdot \bar{\hat{y}}_H}{N_1 \cdot \bar{x}_1 + N_2 \cdot \bar{x}_2 + \cdots + N_{H-1} \cdot \bar{x}_{\text{H-1}} + N_H \cdot \bar{x}_{\text{H}}}$$

The only difference between this expression and the weighted-sum ratio given in the preceding section on stratified ratio estimators is that this expression uses estimated (rather than directly observed) $\hat{y}$ values for the stratum-*H* sample. With this minor adjustment, estimate the population total Y_{all} as:

$$\hat{Y}_{\text{all}}^{(\text{w})} = \hat{r}_{\text{all}}^{(\text{w})} \cdot X_{\text{all}}$$

In these equations, the x_i refer to the *claimed* savings values from the program database (unadjusted) and the X_{all} is the *claimed* total (unadjusted) for the entire population. The standard error is estimated as:

$$\widehat{\mathrm{SE}}\left(\hat{Y}_{\text{all}}^{(\text{w})}\right) = \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right)\sqrt{\sum_{h=1}^{H}\frac{N_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right)\left(s_h^{(r,\,\text{w})}\right)^2 + \left(\frac{N_H}{n_H}\right)^2 \sum_{\text{sample } H} \widehat{\mathrm{SE}}(\hat{y}_i)^2}$$

Here, the standard errors of the $\hat{y}_i$ may reflect adjustments to measure-level *claimed* values, as discussed above.

8.7 Two-Phase (Nested) Sampling

When an M&V protocol requires on-site metering or other labor-intensive procedures at sampled sites, a *two-phase (nested)* design can often reduce study costs without compromising rigor. A two-phase study is conducted as follows:

1. Select a large sample of projects/sites/measures (the Phase 1 sample). Conduct low-cost evaluation research for sites in the Phase 1 sample (for example, phone surveys may be used to verify installation and size or quantity). Use the information obtained to update *claimed* savings values for all sites in the Phase 1 sample.

[46] These claimed adjustments need not be highly detailed, because the final estimate $\hat{y}_i$ will be adjusted to reflect empirical data and rigorous measure-level analysis. The goal is only to reduce $\mathrm{SE}(\hat{y}_i)$ by taking advance measures to diminish the deviations between measure-level verified and claimed savings values.

2. Select a subsample of Phase 1 projects for intensive M&V (this is the Phase 2 sample). Use the M&V data to evaluate verified savings for each of the Phase 2 projects.
3. Analyze the Phase 2 data using a ratio estimator with Phase 1 *claimed* updates as auxiliary data.

In a two-phase study, the total savings is estimated as:

$$\hat{Y} \;=\; \hat{r} \cdot \hat{X} \;=\; \left(\frac{\sum_{\text{Sample 2}} y_i}{\sum_{\text{Sample 2}} x_i}\right) \cdot \left(N \cdot \frac{\sum_{\text{Sample 1}} x_i}{n_1}\right)$$

Because the *claimed* values have been updated to reflect basic verification data, a large source of variation between *claimed* and *evaluated* has been eliminated. This can result in drastic reductions in the effective error ratio. However, the standard error formula needs to be adjusted to reflect the fact that the auxiliary data are only available for a sample and not the whole population. With the adjustment, the standard error is:

$$\widehat{\text{SE}}\big(\hat{Y}\big) \;=\; N \cdot \sqrt{\left(1 - \frac{n_1}{N}\right)\frac{s_y^2}{n_1} + \left(1 - \frac{n_2}{N}\right)\frac{s_{\text{ratio}}^2}{n_2}}$$

Here, s_{ratio} calculated from the deviations between the updated *claimed* values (Phase 1) and the final evaluated savings values (Phase 2).

This approach reconciles two important aspects of evaluation rigor:

- ***Program-level sampling rigor.*** This refers to minimizing sampling error, which is a function of sample size, population size, and variability between reported and verified savings values. (This variability is captured by the error ratio.)
- ***Site-level estimation rigor.*** This refers to minimizing the errors in site-level savings estimates. In other words, minimizing the deviations between a site's verified savings value and its actual savings.

Two-phase sampling may be used to increase sampling efficiency (equivalently, to increase sampling rigor for a given study cost) without reducing site-level evaluation rigor.

Chapter 12: Survey Design and Implementation for Estimating Gross Savings Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

Robert Baumgartner
Tetra Tech
Madison, Wisconsin

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68568
September 2017

Contract No. DE-AC36-08GO28308

Chapter 12: Survey Design and Implementation for Estimating Gross Savings Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

Robert Baumgartner
Tetra Tech
Madison, Wisconsin

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68568
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

These methods, processes, or best practices (“Practices”) are provided by the National Renewable Energy Laboratory (“NREL”), which is operated by the Alliance for Sustainable Energy LLC (“Alliance”) for the U.S. Department of Energy (the “DOE”).

It is recognized that disclosure of these Practices is provided under the following conditions and warnings: (1) these Practices have been prepared for reference purposes only; (2) these Practices consist of or are based on estimates or assumptions made on a best-efforts basis, based upon present expectations; and (3) these Practices were prepared with existing information and are subject to change without notice.

The user understands that DOE/NREL/ALLIANCE are not obligated to provide the user with any support, consulting, training or assistance of any kind with regard to the use of the Practices or to provide the user with any updates, revisions or new versions thereof. DOE, NREL, and ALLIANCE do not guarantee or endorse any results generated by use of the Practices, and user is entirely responsible for the results and any reliance on the results or the Practices in general.

USER AGREES TO INDEMNIFY DOE/NREL/ALLIANCE AND ITS SUBSIDIARIES, AFFILIATES, OFFICERS, AGENTS, AND EMPLOYEES AGAINST ANY CLAIM OR DEMAND, INCLUDING REASONABLE ATTORNEYS' FEES, RELATED TO USER’S USE OF THE PRACTICES. THE PRACTICES ARE PROVIDED BY DOE/NREL/ALLIANCE "AS IS," AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL DOE/NREL/ALLIANCE BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER, INCLUDING BUT NOT LIMITED TO CLAIMS ASSOCIATED WITH THE LOSS OF PROFITS, THAT MAY RESULT FROM AN ACTION IN CONTRACT, NEGLIGENCE OR OTHER TORTIOUS CLAIM THAT ARISES OUT OF OR IN CONNECTION WITH THE ACCESS, USE OR PERFORMANCE OF THE PRACTICES.

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Suggested Citation

Robert Baumgartner. (2017). *Chapter 12: Survey Design and Implementation for Estimating Gross Savings Cross-Cutting Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68568. http://www.nrel.gov/docs/fy17osti/68568.pdf

Acronyms

AAPOR	American Association for Public Opinion Research
CASI	computer-assisted self interviewing
CATI	computer-assisted telephone interviewing
DOE	U.S. Department of Energy
EM&V	evaluation, measurement, and verification
HVAC	heating, ventilating, and air conditioning
RDD	random-digit dialing
SIC	Standard Industrial Classification
TSE	total survey error
UMP	Uniform Methods Project

Table of Contents

1 Introduction

Survey research plays an important role in evaluation, measurement, and verification (EM&V) methods for energy efficiency program evaluations, as the majority of energy efficiency program evaluations use survey data.

EM&V efforts are only as accurate as the data used in analyses. However, despite the prominent role of survey research in EM&V for energy efficiency programs, it is rare to see descriptions of survey research methods and procedures presented in sufficient detail for readers to evaluate the quality of data used in generating the findings.

This chapter presents an overview of best practices for designing and executing survey research to estimate gross energy savings in energy efficiency evaluations. A detailed description of the specific techniques and strategies for designing questions, implementing a survey, and analyzing and reporting the survey procedures and results is beyond the scope of this chapter. So for each topic covered below, readers are encouraged to consult articles and books cited in *References*, as well as other sources that cover the specific topics in greater depth.

This chapter focuses on the use of survey methods to collect data for estimating gross savings from energy efficiency programs. Thus, this section primarily addresses survey methods used to collect data on the following:

- Characteristics of energy consumers (residential and nonresidential), including appliance and equipment ownership and reported behaviors (The results of a well-designed survey help in estimating gross savings attributable to energy efficiency programs.)
- Verification of installation, hours of use, operating conditions, and persistence of new energy-efficient equipment
- Estimation of self-reported changes in behaviors used by households or businesses in response to energy feedback information
- Market characteristics and sales of appliances and equipment (This information is used to establish a baseline for evaluating the impact of energy efficiency programs on market transformation.)
- Estimation of the response to retrofit and energy audit programs designed to increase the efficiency of energy use in households and businesses.

As surveys also provide the primary means of identifying and assessing non-programmatic effects, such as free-ridership, spillover, and market effects, they provide the basis for calculating net savings.

In defining and describing best practices for survey research, the American Statistical Association states (American Statistical Association 1980): “The quality of a survey is best judged not by its size, scope, or prominence, but by how much attention is given to dealing with the many important problems that can arise.” Evaluating survey research and survey data in the manner described in that quotation requires:

- An understanding of the different sources and problems that can arise in designing and executing survey research
- An awareness of best practices for preventing, measuring, and dealing with these potential problems.

This chapter contains guidelines for selecting appropriate survey designs and recommends some administration procedures for different types of energy efficiency EM&V surveys.[1]

[1] As discussed in the section "Considering Resource Constraints" of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

2 The Total Survey Error Framework

Total survey error (TSE) is a framework that allows researchers to make informed decisions for maximizing data quality by minimizing TSE within the constraints of a given research budget (Groves and Lyberg 2010). The TSE framework (widely used as a paradigm in survey research) is applied in evaluating specific types of survey research design. It is also used in evaluating the survey data collected to measure the behaviors of energy consumers for estimating gross savings resulting from energy efficiency programs.

In addition to TSE, other sources of error—such as modeling decisions, low internal and/or external validity, and use of an inappropriate baseline—may also be present in estimates of gross energy savings. However, this chapter deals only with TSE. (Other chapters discuss the appropriate use of modeling and research design for specific end-uses, such as lighting, HVAC, and retrofits.)

For this chapter, the following key terms require definition:

- ***Population of interest.*** The population to which results are to be generalized, sometimes known as the "target" population.
- ***Sampling frame.*** A directory, database, or list covering all members (or as many as possible) of the population of interest.
- ***Sampling element and unit of analysis.*** Persons, groups, or organizations from which data are to be collected.
- ***Survey errors.*** Deviation of a survey response from its underlying true value, caused by random sampling error, coverage error, nonresponse error, and measurement error.
- ***Mode-effects.*** Differences in the same measure, arising from differences in the mode of data collection used (such as interviewer-administered and self-administered surveys).

2.1 TSE Framework for Evaluating Survey and Data Quality

TSE provides a basis for developing a cost-benefit framework by describing statistical properties (or fitness for use) of survey estimates that incorporate a range of different error sources. The development of a cost-benefit framework is beyond the scope of this chapter; however, Groves (Groves 1989) describes how to reduce errors using the principles of TSE in combination with data on the costs of specific survey procedures.

Within a sample of respondents representing the population of interest, TSE recognizes that survey research seeks to measure accurately particular constructs or variables. For a specific survey, resulting measures might deviate from this goal due to four error categories:

- Sampling errors
- Nonresponse errors
- Coverage errors
- Measurement errors.

The TSE framework explicitly considers each of these potential error sources and provides guidelines for making decisions about allocations of available resources. The result is that the sum of these four error sources (the total survey error) can be minimized for estimates developed from survey data.

The subsequent sections contain discussions of each error type and its relevance to EM&V for energy efficiency programs. This chapter also describes current best practices for identifying, measuring, and mitigating these errors.

2.2 Sampling Errors

Sampling errors are random errors resulting from selecting a sample of elements from the population of interest, rather than from conducting a census of the entire population of interest. For practical or monetary reasons, it is often necessary to use a sample relative to an entire population. Although differences will likely occur between the sample and the population, so long as the sample has been based on probability sampling methods, these differences will likely be insubstantial.

A sampling error is the TSE component that is most frequently estimated, using measures such as the standard error of the estimate. Two methods commonly used to reduce sampling error are increasing the sample size or ensuring the sample adequately represents the entire population. (Sample designs, sampling errors, confidence intervals and precision of estimates, and sample selection are discussed in Chapter 11: *Sample Design*)

2.3 Nonresponse Errors

For any survey, some sampled customers likely will not complete the survey. Consequently, nonresponse error may occur if the nonrespondents differ from the respondents on one or more variables of interest. Nonresponse error may also occur when respondents fail to answer individual questions or items in the survey. Note that “nonresponse” is not necessarily the same as “nonresponse bias.” Such bias occurs when differences emerge between respondents and nonrespondents on one or more measures important to the analysis of gross energy savings.

For energy efficiency EM&V surveys, the salience of the topic likely corresponds to the survey response rate (that is, interested individuals are more likely to respond). Consequently, nonresponse bias should be treated as a potential issue in designing survey implementation procedures.

2.3.1 Best Practices for Minimizing Nonresponse Errors

The following techniques have proven effective in reducing nonresponse among various target audiences:

- **Reduce the respondents’ costs in completing surveys.** This is done by building trust and legitimacy in the respondents’ eyes and by convincing the respondents they will receive a benefit from responding. The tools for this include advance letters, follow-up attempts, extending the data collection period, and incentives.
- **Highlight sponsorship of the survey** when it involves an organization with high credibility among the respondents, such as an electric or gas utility, a regulatory

commission, a state or federal agency (for example, the U.S. Department of Energy), or a respected non-governmental organization. Having a credible sponsor usually increases the response rate.

- **When surveying organizations, identify appropriate respondents** to report on an organization's behalf. Then appeal to that individual to respond as the organization's representative. If a superior in the organization identifies an individual as the designated respondent, cite the superior when corresponding with the target respondent.
- **Avoid defining specific survey topics when introducing the survey** to sampled customers. Rather, describe the survey in terms as general as possible to reduce the likelihood of respondents making selections by their interest in a topic.

The potential for nonresponse bias can be estimated using these methods:

- **Collecting data (often a subset of survey questions) from nonrespondents** offers the most direct measure of nonresponse bias, although it can be difficult to obtain a representative sample of nonrespondents.
- **Comparing the responses of early responders (responders on the first contact) with those of responders who are more reluctant or difficult to reach.** This strategy assumes similarities between nonrespondents and reluctant or hard-to-reach respondents.

Where the potential for nonresponse bias has been identified, it is possible to weight the data to attempt to correct for underrepresentation of specific segments of the population. For example, where characteristics of the population are known, sample weights can be developed to adjust the proportion of these characteristics in the sample to match the characteristics of the population. Even when sample weights are used to adjust for nonresponse, however, the researcher has no assurance that the results account for differences between the individual respondents and nonrespondents from a particular segment.

2.4 Coverage Errors

When a sample (even a probability sample) excludes certain members of the population of interest, coverage errors may occur due to differences between the portions of the population excluded and the remainder of the population. A common example of this is a telephone survey that omits households without landlines. This also occurs in surveys of organizations that are selected based on their Standard Industrial Classification (SIC) codes, because new businesses may not have been classified yet and some businesses may have been classified incorrectly. Non-coverage might also result from the exclusion of some population members due to geographic areas, language differences, physical challenges impairing the ability to respond, and individuals living in institutions.

An issue currently faced when using general population telephone surveys is the increasing number of households without landline telephones—recently estimated at more than 30% of all U.S. households (Blumberg and Luke 2011). The likelihood of a household being "wireless only" relates to a number of demographic characteristics, such as:

- Age (younger adults are less likely to have landlines)
- Household types (unrelated adults living together are more likely to be wireless)

- Own/rent status (renters are more likely wireless)
- Household income (adults living in poverty are more likely wireless).

Further, the study indicated that one in six adults in the United States receives most or all telephone calls on wireless phones, even though there is a landline telephone at the residence. These data suggest telephone survey samples that do not include wireless phone numbers may produce data subject to "coverage error." (However, for surveys of program participants in which customers provided contact information, the chance of coverage bias due to missing cell phone-only households is reduced.)

A related issue is the "do not call" list maintained by some utilities. Customers who have requested that they not be contacted regarding certain matters are a potential source of coverage bias for energy efficiency surveys.

2.4.1 Best Practices for Minimizing Coverage Errors

The following techniques have proven effective in reducing nonresponse among various target audiences:

- Evaluate the sample frame carefully to determine whether the listings match populations of interest. In your review, consider these questions: (1) Is the list up to date? (2) Are telephone numbers or other contact information current? (3) Does the list include wireless and landline phone numbers?
- Use dual sampling frames for general population surveys. For example, use cell phone number samples in addition to directory-based (land-line) samples.
- Define the population accurately for which the survey results are appropriately generalized. Thus, any segments not covered in the sample frame are clearly identified.

2.5 Measurement Errors

For most surveys, measurement error presents the most common and problematic error type. The term "measurement error" covers all biases and random variance arising when a survey does not measure its intended target. (This discussion does ***not*** include random errors, where respondents might answer a question differently over repeated trials. That results in increased variance, but not bias.)

In this chapter, measurement error is described as a systematic pattern or direction in differences between respondents' answers to a question and the correct answer. Such error occurs during data collection, rather than from sampling, nonresponse, coverage, or data processing. For example, respondents tend to over-report behaviors they believe are looked upon favorably and underreport behaviors they believe are viewed unfavorably (social desirability bias).

Measurement error results from the following factors:

- Respondent behaviors or responses to questions
- Interviewers' influence on respondents' answers (interviewer effects)
- Question and questionnaire design
- Survey method of administration (mode).

The next sections describe how each of the first three measurement error sources can affect data quality and the best practices for reducing these effects. At the end of this section is a list of best practices for minimizing measurement errors. The effects of survey administration methods on measurement error are discussed in *Survey Administration (Mode) Considerations*.

2.5.1 Respondent Behaviors and Responses

Social desirability, acquiescence bias, and recall errors present the three most relevant bias sources, based on respondent behaviors.

2.5.1.1 Social Desirability Bias

This refers to the tendency of respondents to misreport their attitudes or behaviors intentionally in ways that make them seem appear to be doing "the right thing" in the eyes of interviewers or researchers. For example, in more than 50 years of behavior studies on voting, survey respondents have consistently reported voting at a higher rate than the turnout at the polls has actually indicated. Similarly, as energy efficiency actions are widely viewed as socially desirable behaviors, it is expected that some respondents will over-report that they engaged in energy-efficient behaviors or would have purchased an energy-efficient appliance even had a rebate not been offered.

Voting behaviors provide a common focus for the study of socially desirable responding, as a well-established measure exists (official records of voter turnout) against which voting self-reports can be validated. However, no such validator exits for measures designed to determine whether a respondent would have purchased an energy-efficient appliance without an incentive. Thus, for questions about energy efficiency actions and behaviors, wording that legitimizes socially undesirable behavior can be used to mitigate social desirability bias. (This strategy has also been shown to reduce social desirability bias in surveys of voting behavior.)

For energy efficiency surveys, a question measuring self-reports of energy efficiency actions taken by respondents might be worded as:

> We often find that people have not done things to reduce energy use in their homes. They aren't sure how to do them, they don't have the right tools, or they just haven't had the time. For each of the following activities, please tell me if you have done this in your home. (Holbrook and Krosnick 2010)

Social desirability bias primarily emerges as an issue for interviewer-administered surveys. Consequently, removing the interviewer's presence for self-administered survey modes reduces the pressure for socially desirable responding.

2.5.1.2 Acquiescence Bias

This refers to the tendency for respondents to (1) select an "agree" response more often than a "disagree" response or (2) select a positively-worded response category more often than a negatively-worded response category, regardless of a question's substance.

In several studies using split-sample question wording experiments, Schuman and Presser (1996) demonstrated a classic example of acquiescence bias. They consistently found a difference between the percentage of respondents selecting the "agree" response when asked to agree or disagree with this: "Most men are better suited emotionally for politics than women." This wording received a higher "agree" rate than did the question, "Would you say that most men are better suited emotionally for politics than are most women?"

When respondents were presented with a forced choice question in other response categories indicating that men and women were equally suited or that women were better suited than men in this area, the result was a consistently lower agreement rate. For questions asked in the agree/disagree format, the percentage of responses indicating men were better suited for politics was consistently from 10 to15 percentage points higher than the results of the forced-choice format.

In questions asking about energy efficiency actions, acquiescence bias is expected when statements are worded in a positive direction.

2.5.1.3 Recall Errors

These present another potential bias source based on respondent behaviors. Survey questions often ask respondents to recall specific events or to report on the frequency with which they have engaged in certain behaviors. Cognitive scientists and survey researchers have identified these factors correlating with errors in recall of retrospective events or behaviors:

- **Intervening related events** or new information related to the original event may cause individuals to lose the ability to recall accurately the specific details of any one event.
- **Recall becomes less accurate** with the passage of time.
- **Salient events are remembered more accurately** than less-salient events (Eisenhower et al. 1991). For energy efficiency evaluations, the length of a recall period can be an important element in estimating gross energy savings. Respondents typically are asked to recall whether an event (such as purchase of an energy-efficient appliance) or the frequency of a behavior (such as the number of CFLs purchased) occurred within a specified time period.
- **Recollections of relatively infrequent events, such as purchases of a major appliance, are subject to telescoping errors.** That is, the events may have occurred earlier or later than was reported. Respondents purchasing a major appliance relevant to the survey but outside of the specified timeframe may report the event as occurring within the timeframe.
- **Recall decay—**the inability of respondents to recall events or frequencies of behaviors—tends to affect the accuracy of a respondents' recall of the frequency of relatively routine events (such as the number of CFLs purchased in a specific period).

2.5.2 Satisficing

One way respondents may introduce measurement error into their responses is by "satisficing"—taking actions enabling one to meet the minimum requirements for fulfilling a request or achieving a goal. When a survey question requires a great deal of cognitive work, researchers have found that some respondents use satisficing to reduce that burden (Krosnick 1991). The following behaviors have been observed in respondents attempting to reduce the amount of cognitive effort involved in responding to a survey:

- Choosing "no opinion" response options frequently when it is offered
- Using the same rating for a battery of multiple objects rated on the same scale
- Tending to agree with any assertion, regardless of its content (acquiescence bias)
- Choosing socially desirable responses.

Satisficing tends to occur in questions designed to measure knowledge, attitudes, and self-reports of behavior. The likelihood of respondents' engaging in satisficing is associated with respondents' cognitive abilities, motivations, and task difficulties.

2.5.3 Interviewer Errors and Effects

In interviewer-administered surveys, the interviewer's presence can negatively influence the quality of survey data in several ways, as noted below and in the extensive literature addressing interviewer errors and effects in sample surveys (Biemer et al. 1991):

- As an interview is a social interaction, both the observable characteristics of interviewers and the manner in which interviewers interact with respondents can influence responses to survey questions.
- Interviewers can administer surveys differently to different respondents. For example, interviewers may (1) fail to follow skip patterns correctly, (2) ad lib or change the wording of specific questions, or (3) falsify data.
- In response to respondents' questions or difficulties, interviewers may probe or offer assistance in ways that affect respondents' answers.

The use of telephone interviews and self-administered surveys eliminates some potential effects related to social interactions between interviewers and respondents. Interviewer training—especially training that entails monitoring performance during interviews—provides the most effective way to identify and address potential sources of interviewer errors and effects.

2.5.3.1 Questionnaire and Question Design

Researchers tend to view questionnaires and questions as measurement devices, eliciting information from respondents. As a result, respondents' perspectives are frequently overlooked when questionnaires and questions also serve as a source of information for respondents to draw upon as they provide useful, informative answers to questions asked (Schwartz 1999).

Both the questionnaire (layout, formatting, and length) and the questions (wording, response categories, and context and order of questions) present information to respondents and thus can affect responses.

2.5.3.1.1 Questionnaire Length

It is commonly known that the longer the questionnaire, the more likely it is that respondent fatigue or loss of concentration becomes an issue. However, the answer to the question, "How long is too long?" differs for different survey modes and topics. The interviewer's skill is also a critical factor in terms of developing rapport with a respondent and maintaining the respondent's motivation.

In general, long surveys can be completed most successfully through personal interviews, while telephone surveys are most likely to be completed successfully when they are short. There is less of a consensus on the effect of questionnaire length for self-administered surveys (mail and Internet). Some research suggests that self-administered survey modes, especially Internet surveys, need to be relatively short to prevent respondents from abandoning the survey before it is completed. However, experience has shown that long self-administered surveys (ranging from 20 to 30 minutes) can be successfully administered, especially for mail questionnaires.

2.5.3.1.2 Open-Ended and Closed-Ended Questions

Although the great majority of energy efficiency evaluation survey questions are closed-ended, there are advantages to using an open-ended format for certain questions. For example, some researchers believe that open-ended questions about quantities—such as the numbers of times a respondent visited a specific website—produce less bias than closed-ended questions. Specifically, this tends to apply to grouped, closed-ended response categories, such as "at least one time per week" and "one to three times per month."

Response categories for closed-ended questions convey information about researchers' expectations. Also, many respondents tend to avoid extreme (high and low) scale points. However, an open-ended question for which response categories are not provided avoids potential data-quality issues.

Similarly, for questions addressing the relative importance of issues facing the country, the closed-ended response categories offered to respondents indicate the issues that researchers think are most likely to be mentioned. This reduces the likelihood of respondents addressing issues not on the list. Despite this, closed-ended questions are used more often, as they are easier to code, process, and analyze. A general rule for using closed-ended questions is to ensure the response categories are comprehensive (Krosnick and Presser 2009).

2.5.3.1.3 Respondents' Interpretation of Questions

Because respondents must understand questions being asked, the researcher must determine whether the respondents' understanding of the questions matches the researcher's intent. Even for a seemingly straightforward question (for example, "What things do you typically do in your household every day to conserve energy?"), it is important to have some knowledge of the respondents' typical tasks.

Differences tend to occur in the literal understanding of the question (Schwartz 1999). For example, although respondents are likely to understand the literal meaning of a question, they must still determine the types of actions or activities of interest to the researcher. Consequently, in surveys about energy efficiency, respondents may ask themselves questions such as:

- "Should I report turning off lights when I leave the room, or is that too obvious?"
- "If I have an automatic set-back thermostat, is that considered an everyday activity?"

For questions open to multiple literal interpretations, researchers can guide respondents by using common examples of the types of information sought.

2.5.3.1.4 Question Order

The order of questions in a survey affects responses. When answering a specific question, respondents are likely influenced by cues and information from previous questions. For example, previous questions can present a priming effect—making certain issues more salient. Asking about the importance of energy efficiency before asking respondents about their energy efficiency behaviors likely implies that those behaviors should be consistent with respondents' stated views on the importance of energy efficiency.

2.5.4 Best Practices for Minimizing Measurement Errors

- **Use pretesting to identify potential measurement errors**, such as instances in which respondents either misinterpret a question or are unable to provide an accurate answer.
- **Use salient events or dates in recall questions** to mark the relevant time period (bounded recall). Where possible, reduce burdens on respondents by shortening the recall periods.
- **Word the questions carefully** so respondents understand it is permissible to report engaging in non-socially desirable behaviors.
- **Use cognitive interviewing** as part of the survey pretest to explore how respondents interpret the questions and construct responses (Madans et al. 2011).
- **To minimize acquiescence bias, avoid "agree/disagree" questions.** Instead, use questions explicitly presenting positive (agree) and negative (disagree) responses in the question stem, such as: "Would you say that most men are better suited emotionally for politics than are most women, that men and women are equally suited, or that women are better suited than men in this area?"
- **Use multiple-item measurement scales when assessing attitudes or reported behaviors**, and pre-test these scales to ensure unidimensionality and internal consistency. A multiple-item measurement scale consists of a number of individual questions combined into a single value. Using multiple-item measures usually increases the reliability of the measure.
- **Train interviewers and monitor the quality of their work** through observational interviews to reduce interviewer errors and interviewer effects.

2.5.5 Best Practices for Measuring Self-Reports of Behaviors

Evaluations of energy efficiency programs often use self-reports of energy-efficient behaviors (or behavioral intentions). Thus, self-report surveys are designed to (1) identify barriers in achieving gross energy savings and (2) help explain differences in energy consumption between treatment and control group customers in programs with experimental designs. The best practices for these surveys of attitudes, behaviors, and behavioral intentions are described in the following sections.

2.5.5.1 Multiple Item Measurement Scales

Since the 1930s, survey researchers have used multiple-item scales to measure attitudes or reported behaviors. Based in psychometric theory, the rationale for multiple-item, self-reported behavior measurement suggests four primary advantages:

1. A set of multiple items can represent the construct (attitude or behavioral report) more completely than can a single item.
2. Combining items reduces potentially idiosyncratic influences of any single item.
3. Aggregating across items increases the reliability (or precision) of measures.
4. Using multiple items more finely distinguishes among respondents, potentially providing a measurement scale appropriately treated as continuous (Nunnally 1978).

In many cases, multiple-item scales of attitudes or self-reported behaviors treated as interval-level or continuous variables (item 4 in the list above) present important implications for statistical analyses of these data. Measures of central tendencies or dispersions prove appropriate for interval or continuous variables, and relative differences in scores between groups of respondents can be calculated. Multiple-item scales also produce variables well suited for use in regression models estimating gross energy savings.

Two procedures have allowed the development of summated multiple-item measures:

1. Factor analysis to verify multiple items measuring a single underlying construct (unidimensionality)
2. A measure of internal consistency using Cronbach’s alpha (coefficient of reliability) or a similar measure of the internal consistency of the measurement scale.

3 Developing Questions

To measure respondent self-reports of attitudes or behaviors in closed-ended questions, the design of the questions entails decisions about these critical elements:

- The order of response categories to be presented to respondents
- The use of a rating or ranking scale
- The type of rating scale
- The use of a middle or neutral category in a rating scale.

A summary of current evidence and best practices for each of these decisions is discussed below.

3.1 Order of Response Alternatives

The responses to closed-ended questions can be influenced by the order in which response categories are presented. For self-administered questionnaires and "show cards" used in personal interviews—where response categories are presented visually—research has shown a primacy effect often occurs. That is, respondents tend to select the answers offered early in the list. However, where response categories are presented verbally by an interviewer (whether on telephone or in person), a recency effect tends to occur, where respondents select answers offered later in the list (Sudman et al. 1996). These research findings demonstrate the need to rotate the order of response alternatives offered to respondents.

3.2 Rating or Ranking?

Although rating scales commonly are used in energy efficiency evaluation surveys, some situations have shown ranking to be a more effective method for measuring the importance of a specific issue or behavior. When the primary goal for a question is to determine the order of two or more objects, a ranking format may be most useful (Visser et al. 2000).

3.2.1 Use of Ranking Scales

Ranking scales avoid the problems of non-differentiation, which occur when rating scales produce very similar ratings for a set of objects. However, rating scales are more commonly used in energy efficiency evaluation surveys for the following reasons:

- Ranking is a more cognitively difficult task for respondents to complete, especially when dealing with a relatively large number of items
- Ranking scores prove more difficult to analyze. (As no assurance exists of equal distances between rankings, they cannot be used appropriately as interval measures.)

3.2.2 Use of Rating Scales

As previously mentioned, rating scales are the predominant method used for measuring self-reports of attitudes or behaviors. The basic types of these scales are classified as:

- Bipolar (from negative to positive, with a neutral point in the middle)
- Unipolar (from a zero point to a highly positive point, such as a range from "no importance" to "extremely important").

After selecting the type of rating scale to use, the next decision is the length or the number of points on the scale. A quick review of questionnaires for energy efficiency evaluations yields a wide range, from dichotomous (yes/no) scales to scales having as many as 100 points.

An important consideration in such decisions is whether to use scale points that divide the continuum into equal distances. If, for example, a scale offers a choice between "poor," "good," and "very good" but these choices have no numeric labels, then the continuum is not divided equally, as "good" and "very good" appear more closely related than "good" and "poor."

Scales using numerical labels meet the "equal interval" requirement. Many studies suggest data quality can be improved by labeling all scale points, rather than labeling only end points and neutral points (Krosnick et al. 1999). Study findings indicate that applying these two techniques improves the results:

- Using words to anchor end-points and perhaps mid-points
- Using numbers to label each point on the scale.

As to the optimal number of scale points, reviews of research show the greatest measurement reliability results from seven-point scales for bipolar scales and five-point scales for unipolar scales.

3.2.3 Use of Middle Alternatives or Neutral Scale Points

Having a middle alternative (or a neutral alternative) increases the reliability of a measure, according to studies that examined the differences in reliability of an item's measurement —specifically, the use of a middle alternative in a scale (O'Muircheartaigh et al. 1999). Some researchers advise using a middle category in a rating scale when a significant number of respondents are likely either to be uninformed or to have no opinion on the issue. Research also shows that the use of a middle alternative changes the frequency distribution of responses across all categories, but it often does not affect the ratio of responses on either side of the scales' middle point (Schuman and Presser 1981).

A recent alternative is to omit the middle category and then measure the intensity of the attitude. In this option, using a scale ranging from "strongly agree" to "strongly disagree" enables researchers to separate those who definitely hold a certain attitude from those who are simply inclined in a particular direction (Converse and Presser 1986). A number of experimental studies have shown data quality for a specific measure usually does not differ significantly, regardless of whether a neutral/no-opinion scale point is offered (Schuman and Presser 1996). In a 2002 study, Krosnick reported:

> The vast majority of neutral or no-opinion responses are not due to completely lacking an attitude, but are most likely to result from a decision not to do the cognitive work necessary to report it (satisficing), a decision not to reveal a potentially embarrassing attitude (social desirability bias), ambivalence, or question ambiguity.

This suggests the best practice for measuring attitudes or behavioral intentions entails omitting the neutral or no-opinion response category and encouraging respondents to report whatever opinion they have.

3.3 Summary of Best Practices for Question Design and Order in a Questionnaire

In their chapter on the design of questions and questionnaires, Krosnick and Presser advise the following when designing survey questions (Krosnick and Presser 2009):

- Use simple, familiar words, avoiding jargon, technical terms, and slang.
- Avoid words with ambiguous meanings; aim for words that all respondents interpret the same way.
- Use specific and concrete wording rather than general and abstract terms.
- Make response categories exhaustive and mutually exclusive.
- Avoid leading or loaded questions that push respondents toward an answer.
- Ask one thing at a time; avoid double-barreled questions.
- Avoid questions with single or double negations.

Further, Krosnick and Presser offer this advice regarding question order:

- To build rapport between respondents and researchers, make early questions easy and pleasant to answer.
- Questions at the beginning of a questionnaire should explicitly address the survey topic, as described to the respondent before the interview.
- Questions on the same topic should be grouped together.
- Questions on the same topic should proceed from the general to the specific.
- Questions on sensitive topics, which might make respondents uncomfortable, should be placed at the end of the questionnaire.
- Use filter questions to avoid asking respondents questions that do not apply to them.

3.4 Survey Administration (Mode) Considerations

The wide range of data collection modes available to survey researchers tend to fall into one of these categories:

- Interviewer-administered modes, such as personal or face-to-face interviews and telephone interviews
- Self-administered modes, such as mail or Internet surveys.

With advances in information and communication technologies, variations exist for each of the primary data collection modes. For example:

- Personal interviews can be conducted by an interviewer who records responses directly onto a laptop or electronic tablet.
- Self-administered questionnaires can be administered by audio-CASI [computer-assisted self interviewing], with questions recorded on an electronic device and played back to respondents, who enter responses electronically.
- Telephone interviews can be conducted by Webcam, in which respondents use either a voice-over Internet protocol or their phone keys to specify their answers.

The choices of data collection modes for energy efficiency evaluations typically involve assessing strengths and weaknesses of a range of factors such as:

- Ability to access to a representative sample of the population of interest
- Types of questions to be asked
- Cost and time required for implementation
- Length, complexity, and content of the questionnaire.

3.4.1 Face-to-Face Personal Interviews

Considered by many survey researchers to be the "gold standard," face-to-face personal interviews generally result in high response rates, even for relatively long questionnaires (45 minutes or more). Through this approach, interviewers can manage complex questionnaires and those requiring visual or verbal background or explanations for the survey questions. However, face-to-face personal interview surveys are fielded less often due to their relatively high cost, as compared to other survey modes. Other key drawbacks are:

- The longer time required to complete data collection
- The logistical difficulty of quality control measures, such as observing interviewers conducting the interviews
- The potential for interviewer effects resulting from interviewer-respondent interactions.

3.4.2 Telephone Interviews

Telephone interviews have surpassed face-to-face personal interviews as the most common interviewer-administered survey mode for these reasons:

- The relatively lower cost per completed interview
- The availability of off-the-shelf random-digit dialing (RDD) samples of the general population;
- The shorter length of time required to complete data collection; and
- The high proportion of households in the United States with a telephone.

With the advent of computer-assisted telephone interviewing (CATI), telephone surveys can accommodate complex questionnaires that apply skip patterns customized to respondent answers. Also, these interviews can be centrally monitored for quality control.

The key drawbacks of telephone interviews are:

- The comparatively low (and declining) response rates
- The relatively short time respondents can be expected to remain engaged (usually no more than 15 to 20 minutes)
- The increasing number of households using call-screening devices
- The increasing number of households without landline telephones.

Additionally, it is difficult to ask sensitive questions through telephone interviews, and social desirability bias presents a potential threat.

As a result of decreased coverage and response rates, telephone surveys are becoming less representative of the population of interest, except when mobile phone numbers are included in the survey. However, using listed samples of utility customers or program participants who have provided contact information can facilitate the contact of general-population households.

Note that when contacting a respondent by cell phone to conduct a survey, it is strongly recommended that the survey not be conducted if the respondent is driving a motor vehicle at the time of the call. In these cases, the interviewer should be instructed to make an appointment for a better time to call the respondent.

3.4.3 Mail Questionnaire Surveys

While the advantages of having an interviewer administer the questionnaire are noted above, there are also potential advantages for mail and self-administered questionnaires (without an interviewer). Self-administered questionnaires have been shown to (1) produce more accurate or candid data for sensitive questions and (2) reduce social desirability bias.

Mail questionnaires can be sent to anyone with an address. Also, respondents do not have to be home at any specific time, as is required for face-to-face personal interviews or telephone interviews. While completing a mail questionnaire survey, respondents can look up personal records, utility billing statements, or purchase information.

Although mail questionnaires often are described as the lowest-cost alternative among survey modes, this approach—in our experience—requires at least two follow-up mailings and, in some cases, relies on an incentive to increase the response rate. This increases cost of fielding the survey. Other drawbacks typically associated with mail questionnaire surveys are:

- Relatively low response rates (in many cases, rate comparable to a telephone survey)
- Longer data collection periods
- Skip patterns must be relatively simple to avoid confusing respondents
- Loss of control over who answers the questions
- Loss of control regarding the order in which questions are viewed and answered.

3.4.4 Internet Surveys

Internet surveys have increased in popularity, especially as the percentage of households and individuals with access to the Internet has increased. These surveys offer the advantage of lower cost (no expenses for paper, printing, mailing, telephones, or interviewers). Further, once the fixed costs of programming and set-up have been incurred, a much larger sample size can be used―even internationally―with very small marginal cost increases.

Internet surveys usually require very short data collection times, with most responses received within one week, although follow-up contacts should be made with nonrespondents to increase response rates. Note, however, that coverage bias for potential respondents who do not have access to the Internet remains an issue with online surveys.

Consistency in the appearance of the survey is also an issue. While enhanced Internet survey software allows for complex skip patterns and sophisticated graphics, different hardware and software used by respondents can result in differences in a questionnaire's appearance and presentation.

As with mail questionnaire surveys, the absence of an interviewer requires that the questions be relatively simple and straightforward. Still, with Internet surveys, the respondents' willingness to answer sensitive questions candidly is increased and the likelihood of social desirability bias is decreased.

3.5 Using Multiple Survey Modes: Mixed-Mode Surveys

In this century, a major trend in survey research has been the increased use of combined survey implementation modes (Dillman et al. 2009). It has long been a practice to mix modes in:

- The survey's contact phase (for example, using an advance letter to contact respondents for telephone surveys or face-to-face interviews)
- Completing different portions of a survey.

What has been relatively new in survey research, however, is use of mixed-mode surveys in which some respondents provide data using one mode, while others provide data using a second (or third) mode (Couper 2011).

This section describes this relatively new approach to mixed-mode surveys. Their increasing use has been driven by several factors, including declining response rates, coverage problems in single-mode surveys, and the development of new survey modes—such as interactive voice response (IVR) and Internet-based methods.

Research has shown that mixed-mode surveys can achieve higher response rates and better coverage of populations of interest. As different methods have different strengths and weaknesses, using a variety of methods can provide complementary results (de Leeuw 2005). Still, mixed-mode surveys present drawbacks—such as increased measurement error—because different survey modes can produce different responses to the same question (Christian et al. 2008).

In a 2011 publication addressing questions about using a mixed-mode survey, Mick Couper cited two strategies in dealing with potential mode differences:

- The **unimode construction** approach constructs questionnaires to be as identical as possible.
- The **correction approach** entails accepting fundamental differences in data collection by different modes *and* designing the data collection instrument to maximize the benefits of each mode; statistical adjustments then are made across the modes used. (Couper 2011.)

A third strategy is to combine these approaches when designing and implementing mixed-mode energy efficiency evaluation surveys. For example, in mixed-mode surveys using telephone and Internet, the fixed-page telephone interview survey—where respondents are asked questions in a specified sequence by CATI—can best be replicated by an Internet survey, where respondents see one question at a time, and cannot progress to the next question until the first is answered. Also, an IVR Internet survey can also be used to replicate the presence of an interviewer for such mixed-mode surveys.

For a mixed-mode survey using mail and Internet questionnaires, the scrolling-page Internet survey design best replicates mail questionnaire design, where respondents can turn ahead pages if they wish to see questions in the survey.

Replicating in two survey modes how questions are presented provides an opportunity to increase the effectiveness of energy efficiency evaluation surveys, while increasing coverage and response rates. New technologies and advancements in survey research capabilities will continue to provide additional ways of mixing modes and to increase survey effectiveness and quality.

4 Minimum Reporting Requirements for Energy Efficiency Evaluation Surveys

Survey research organizations—such as the American Association for Public Opinion Research (AAPOR) and the Council of American Survey Research Organizations—require their members follow appropriate professional guidelines for disclosing and reporting survey methods and findings. The goal of these organizations is to advance the state of knowledge and practice by providing sufficient information to permit review and replication by other researchers.

AAPOR offers various guidelines regarding the minimum essential information on survey methods to be disclosed in research reports:

- Survey sponsor and the firm conducting the survey
- Survey purpose and specific objectives
- Questionnaire and exact/full wording of questions as well as any other instructions or visual exhibits provided to respondents
- Definitions of populations under study
- Descriptions of the sampling frame used to identify populations under study
- Sample design, including clustering, eligibility criteria and screening procedures, selection of sample elements, mode of data collection, and the number of follow-up attempts
- Sample selection procedures (how sample cases were selected)
- Documentation of response or completion rates, numbers of refusals, and other dispositions
- Discussion of the findings' precision, including sampling error, where appropriate
- Descriptions of special scoring, editing, data adjustment, or indexing procedures used
- Methods, locations, and dates of fieldwork or data collection
- Copies of interviewer instructions for administering the questions.

Following the disclosure and reporting guidelines available on the AAPOR website serves to advance knowledge and the state of practice for energy efficiency evaluation research and, ultimately, results in better-quality data and better decisions on energy efficiency programs.

5 Conclusion

This chapter has provided an overview of the current state of survey research regarding the evaluation of energy efficiency programs through (1) developing estimates of gross energy savings, (2) determining well market effects, and (3) identifying process issues. For each topic covered—summarized below—readers are encouraged to consult articles and books cited in *References* as well as other sources covering these topics in much greater depth:

- Sources of survey error, such as nonresponse, coverage, and measurement
- Best practices for measuring self-reports of attitudes and behaviors
- Best practices for question wording and question order
- Selection of survey modes and use of mixed-mode approaches
- Minimum guidelines for reporting and disclosure of survey research.

6 References

American Association for Public Opinion Research (AAPOR). www.aapor.org/Best_Practices1.htm.

American Statistical Association. (1980). *What Is a Survey?* Washington, DC.

Biemer, P.; Groves, R.M.; Lyberg, L.E.; Mathiowetz, N.A.; Sudman, S., eds. (1991). *Measurement Errors in Surveys*. John Wiley & Sons.

Blumberg, S.J.; Luke, J.V. (2011). "Wireless Substitution: Early Release of Estimates From the National Health Interview Survey." Division of Health Interview Statistics, National Center for Health Statistics, Centers for Disease Control. www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201112.pdf.

Christian, L.M.; Dillman, D.A.; Smyth, J.D. (2008). "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys." Lepkowski, J.; Tucker, C.; Brick, M.; de Leeuw, E.D.; Japec, L.; Lavrakas, P.; Link, M.; Sangster, R. eds. *Advances in Telephone Survey Methodology*. Wiley-Interscience. www.sesrc.wsu.edu/dillman/papers/2006/theeffectsofmodeandformat.pdf.

Converse, J.M.; Presser, S. (1986). *Survey Questions: Handcrafting the Standardized Questionnaire.* Sage Publications.

Couper, M.P. (2011). "The Future of Modes of Data Collection." *Public Opinion Quarterly.* (75:5); pp. 889-908. http://poq.oxfordjournals.org/content/75/5/889.full.pdf+html.

de Leeuw, E.D. (2005). "To Mix or Not to Mix Data Collection Modes in Surveys." *Journal of Official Statistic*s. (21:2); pp. 233-255. http://igitur-archive.library.uu.nl/fss/2011-0314-200305/EdL-to%20mix%202005.pdf.

Dillman, D.A.; Phelps, G.; Tortora, R.; Swift, K.; Kohrell, J.; Berck, J.; Messer, B. (2009). "Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR), and the Internet." *Social Science Research*. (38:1); pp. 1-18.

Eisenhower, D.; Mathiowetz, N.A.; Moganstein, D. (1991). "Recall Error: Sources and Bias Reduction Techniques." Biemer, P.; Groves, R.M.; Lyberg, L.E.; Mathiowetz, N.A.; Sudman, S. eds. *Measurement Errors in Surveys*. John Wiley & Sons. http://onlinelibrary.wiley.com/doi/10.1002/9781118150382.ch8/summary.

Groves, R.M.; Lyberg, L. (2010). "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly*. (74:5); pp. 849-879. http://poq.oxfordjournals.org/content/74/5/849.full.pdf.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. Wiley Series in Survey Methodology. Wiley-Interscience: New York.

Holbrook, A.L.; Krosnick, J.A. (2010). "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item-Count Technique." *Public Opinion Quarterly.* (74:2); pp. 37-67. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1569295##.

Krosnick, J.A. (1991). "Response Strategies for Coping with Cognitive Demands of Attitude Measurement in Surveys." *Applied Cognitive Psychology.* (5); pp. 213-236.

Krosnick, J.A.; Presser, S. (2009). "Question and Questionnaire Design." Wright, J.D.; Marsden, P.V. eds. *Handbook of Survey Research (2nd Edition).* Elsevier: San Diego. http://comm.stanford.edu/faculty/krosnick/docs/2010/2010%20Handbook%20of%20Survey%20Research.pdf.

Krosnick, J.A.; Robinson, J.P.; Shaver, P.R; Wrightsman, L. eds. (1999). "Maximizing Questionnaire Quality." *Measures of Political Attitudes*. Academic Press, San Diego. http://comm.stanford.edu/faculty/krosnick/docs/1999/1999_robinson02_krosnick.pdf.

Madans, J.; Miller, K.; Maitland, A.; Willis, G. (2011). *Question Evaluation Methods: Contributing to the Science of Data Quality.* John Wiley and Sons.

Nunnally, J.C. (1978). *Psychometric Theory* (2nd ed.) McGraw-Hill, New York.

O'Muircheartaigh, C.; Krosnick, J.; Helic, A. (1999). "Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data." Paper presented at the American Association for Public Opinion Research Annual Meeting, St. Petersburg, Florida. http://ideas.repec.org/p/har/wpaper/0103.html.

Schuman, H.; Presser, S. (1996). *Questions and Answers in Attitude Surveys*: *Experiments on Question Form, Wording, and Context*. Sage Publications.

Schuman, H.; Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context.* Academic Press.

Schwartz, N. (1999). "Self-Reports: How the Questions Shape the Answers." *American Psychologist*. (54:2); pp. 93-105. http://psycnet.apa.org/index.cfm?fa=search.displayRecord&uid=1999-00297-001.

Sudman, S.; Bradburn, N.; Schwartz, N. (1996). *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco.

Visser, P.S.; Krosnick, J.A.; Lavrakas, P.J. (2000). "Survey Research." Reis, H.T.; Judd, C.M. eds. *Handbook of Research Methods in Social and Personality Psychology.* Cambridge University Press.

7 Resources

Bradburn, N.; Sudman, S.; Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design–For Market Research, Political Polls, and Social and Health Questionnaires.* John Wiley & Sons.

Wikman, A.; Warneryd, B. (1988). "Measurement Errors in Survey Questions." *Social Indicators Research*. (22:2); pp. 199-212. www.jstor.org/discover/10.2307/27520814?uid=3739568&uid=2129&uid=2&uid=70&uid=4&uid=3739256&sid=21101416225633.

Chapter 13: Assessing Persistence and Other Evaluation Issues Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

Daniel M. Violette
Navigant
Boulder, Colorado

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68569
September 2017

Contract No. DE-AC36-08GO28308

Chapter 13: Assessing Persistence and Other Evaluation Issues Cross-Cutting Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in April 2013.

Daniel M. Violette
Navigant
Boulder, Colorado

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68569
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

These methods, processes, or best practices ("Practices") are provided by the National Renewable Energy Laboratory ("NREL"), which is operated by the Alliance for Sustainable Energy LLC ("Alliance") for the U.S. Department of Energy (the "DOE").

It is recognized that disclosure of these Practices is provided under the following conditions and warnings: (1) these Practices have been prepared for reference purposes only; (2) these Practices consist of or are based on estimates or assumptions made on a best-efforts basis, based upon present expectations; and (3) these Practices were prepared with existing information and are subject to change without notice.

The user understands that DOE/NREL/ALLIANCE are not obligated to provide the user with any support, consulting, training or assistance of any kind with regard to the use of the Practices or to provide the user with any updates, revisions or new versions thereof. DOE, NREL, and ALLIANCE do not guarantee or endorse any results generated by use of the Practices, and user is entirely responsible for the results and any reliance on the results or the Practices in general.

USER AGREES TO INDEMNIFY DOE/NREL/ALLIANCE AND ITS SUBSIDIARIES, AFFILIATES, OFFICERS, AGENTS, AND EMPLOYEES AGAINST ANY CLAIM OR DEMAND, INCLUDING REASONABLE ATTORNEYS' FEES, RELATED TO USER'S USE OF THE PRACTICES. THE PRACTICES ARE PROVIDED BY DOE/NREL/ALLIANCE "AS IS," AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL DOE/NREL/ALLIANCE BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER, INCLUDING BUT NOT LIMITED TO CLAIMS ASSOCIATED WITH THE LOSS OF PROFITS, THAT MAY RESULT FROM AN ACTION IN CONTRACT, NEGLIGENCE OR OTHER TORTIOUS CLAIM THAT ARISES OUT OF OR IN CONNECTION WITH THE ACCESS, USE OR PERFORMANCE OF THE PRACTICES.

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter author wishes to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Hossein Haeri of The Cadmus Group
- Brent Barkett of Navigant.

In addition, the chapter author wishes to acknowledge helpful comments submitted through the Stakeholder Review process.

Suggested Citation

Daniel M Violette. (2017). *Chapter 13: Assessing Persistence and Other Evaluation Issues Cross-Cutting Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68569. http://www.nrel.gov/docs/fy17osti/68569.pdf

Acronyms

AC	air-conditioning
Btu	British thermal unit
CALMAC	California Measurement Advisory Council
CFL	compact fluorescent lamp
DOE	U.S. Department of Energy
EM&V	evaluation, measurement, and verification
EMS	energy management system
EUL	effective useful life
FCM	forward capacity market
HVAC	heating, ventilating, and air conditioning
ISO-NE	Independent System Operator-New England
M&V	measurement and verification
NPV	net present value
RUL	remaining useful life
SAS	Statistical Analytics Software
TDF	technical degradation factor
UMP	Uniform Methods Project

Table of Contents

List of Figures

List of Tables

1 Introduction

Addressing other evaluation issues that have been raised in the context of energy efficiency programs, this chapter focuses on methods used to address the persistence of energy savings, which is an important input to the benefit/cost analysis of energy efficiency programs and portfolios. In addition to discussing "persistence" (which refers to the stream of benefits over time from an energy efficiency measure or program), this chapter provides a summary treatment of these issues:

- Synergies across programs
- Rebound
- Dual baselines
- Errors in variables (the measurement and/or accuracy of input variables to the evaluation).

This first section of this chapter contains a definition of persistence and identifies issues in its evaluation. The state of the practice in persistence is addressed, examples taken from persistence studies are presented, and recommendations for addressing persistence are presented at the end of the section. The other evaluation issues are addressed in the second section of the chapter. Appendix A presents a matrix of persistence issues and methods by program type.[1]

[1] As discussed in the section "Considering Resource Constraints" of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

2 Persistence of Energy Savings

Understanding persistence is critical to making good decisions regarding energy efficiency investments, so this section outlines program evaluation methods that can be employed to assess persistence—the reliability of savings over time. Energy efficiency program benefits are measured as the net present value (NPV) of a stream of benefits based on the energy and demand savings[2] achieved by the program. Depending on the mix of measures and their assumed lives, these benefits may extend to 15 years (or more) for some measures. As a result, assumptions about the persistence of savings over time influence the energy efficiency benefit-cost tests. Extrapolating savings beyond the evaluation period has often been based on engineering judgment, manufacturer specifications, and some empirical work (the factors used to develop projections of measure lifetimes and degradation).

The protocols developed under the Uniform Methods Project (UMP) in other chapters generally focus on estimating first-year savings. There is also some discussion, however, about estimating first- and second-year savings when more participants from a second program year are needed for the impact evaluation. These initial evaluations are often quite detailed, assessing both the savings and the quality of the program in terms of installation, engineering calculations, and equipment selection (where on-site visits are used to validate initial "claimed" estimates).

2.1 Addressing Persistence

Persistence of savings encompasses both the retention and the performance degradation of measures. Together, these factors are used to estimate how the *claimed* persistence values used in program planning can be updated based on *evaluated* savings values.[3] Different jurisdictions define and treat the components of overall persistence differently. As a result, defining what is meant by overall persistence and addressing some of the subtle context issues are important to the discussion.

There are a number of subtle aspects to the context and definition of overall persistence. Consistent and practical definitions for use in developing estimates of the overall persistence of savings over time were developed for the Joint Massachusetts Utilities (Energy and Resource Solutions 2005).[4] In that study, overall persistence is divided into two components: (1) measure life and (2) savings persistence.

Recognizing that definitions for *persistence* and *realization of savings* are not nationally consistent, the definitions based on the Massachusetts framework and outlined below provide a structure that can be addressed by evaluation and verification methods. That is, these definitions

---

[2] This chapter focuses on estimating energy savings, but the persistence of reductions in demand may also be important for some measures and programs. Issues raised here may also be important for programs and policies focused on reducing demand during peak periods.

[3] In this chapter and consistent with other chapters, *claimed* savings means the same as *ex ante* savings and *evaluated* savings is used instead of *ex post* savings. This note is to eliminate confusion for those more familiar with the use of "*ex ante"* (initial savings estimates) and "*ex post"* (evaluated savings) terminology in describing evaluation methods.

[4] This study for the Joint Massachusetts Utilities' defines "measure life" as the median number of years that a measure is installed and operational. This definition implicitly includes equipment life and *measure* persistence. However, *savings* persistence is the percentage of change in expected savings due to changed operating hours, changed process operation, and/or degradation in equipment efficiency relative to the baseline efficiency option.

use categories of effects and factors that can be quantified using evaluation methods. For example, it is difficult to estimate technical measure life based on on-site inspections, as there may be many reasons that a measure is no longer in place. Thus, technical measure life and other reasons for measure non-retention are combined in the definition “measure life,” which is simply the time a measure can be expected to be in place and operable.

2.1.1 Definitions

The definitions of key terms used in this chapter are these.

2.1.1.1 Measure Life or Effective Useful Life

This is the median number of years that a measure is in place and operational after installation. This definition implicitly includes equipment life and measure persistence (defined below), but not savings persistence.

- “Equipment life” is the number of years installed equipment will operate before it fails.
- “Measure persistence” takes into account business turnover, early retirement or failure of the installed equipment, and any other reason the measure would be removed or discontinued.

2.1.1.2 Savings Persistence

This is the percentage of change in expected savings due to changed operating hours, changed process operations, and/or the performance degradation of equipment efficiency relative to the baseline efficiency option. For example, an industrial plant that reduces operation from two shifts to one shift may then have a savings persistence factor of 50%, as only half of the projected energy savings would be realized. Also, improper operation of the equipment may negatively affect *savings* persistence, so training and commissioning could improve savings persistence. Finally, most equipment efficiency degrades over time, so annual energy savings may increase or decrease relative to the efficiency degradation of the baseline efficiency option.

Figure 1 illustrates how the two persistence factors are used to produce savings that are adjusted for persistence: Savings Adjusted for Persistence = (Measure Life Factor) x (Savings Persistence Factor) x (Initial Savings Estimate).

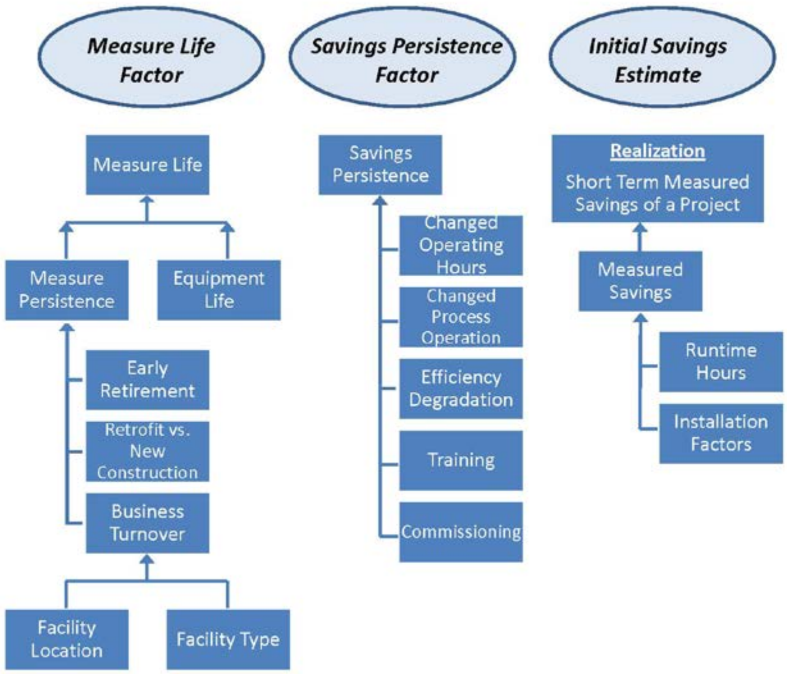


Figure 1. Relationship of measure life, savings persistence, and initial savings estimates[5]

2.1.2 Factors for Selecting a Persistence Study

The following are several important factors to consider when selecting the type of study to examine energy savings persistence.

2.1.2.1 Available Claimed Estimates of Persistence

There are almost always initial *claimed* estimates of the assumed stream of savings for a program (based on current estimates of measure life and degradation). These estimates are used in the initial benefit/cost analyses conducted as part of program design or in the benefit/cost tests of initial program evaluations efforts. As a result, most studies of persistence test the initial claimed stream of savings against the evaluated results to check for significant differences.[6] The outcome is often presented as a realization rate (that is, the *evaluated* values divided by the initial claimed values), which is the year-by-year savings estimate used in benefit/cost studies.

[5] Source: Adapted from Energy and Resource Solutions (2005).

[6] Starting with a set of claimed savings allows for the use of evaluation methods that leverage these initial data through the use of ratio estimates and a "realization rate" framework.

2.1.2.2 Uncertainty in Claimed Estimates

When deciding whether to conduct a new study of persistence—and the corresponding level of effort required—consider the confidence that the evaluator or decision-maker has in the claimed stream of savings values. If the uncertainty is perceived as being high *and* a sensitivity analysis shows that plausible revisions to persistence of energy savings substantively changes the results of benefit/cost tests, then a new study may be worthwhile. Such an undertaking regarding persistence may result in revisions to the current *claimed* estimates.

For example, measures that account for greater savings, have shorter measure life values, or may be subject to near-term degradation in savings are more important to evaluate, as they will have a greater impact on the resulting benefit/cost tests. However, changes in measure life that do not take effect until the 14th or 15th year of the measure may be discounted in the NPV calculation (discussed below). Thus, in terms of the effect on the benefit/cost calculation, the additional work needed to estimate these values may not be worthwhile.

2.1.2.3 Discounting Values of Energy Savings Over the Life of the Measure

The stream of program benefits over time is discounted, resulting in near-term savings estimates that have a larger impact on the NPV of benefits than the values further out in the future. For example, the effect of research on the measure life of a second refrigerator retirement that extends it from six years to eight years would be muted somewhat in the benefit/cost analysis due to discounting. Specifically, the energy savings from this updated measure life of two additional years would be muted in its application by discounting the benefits for year seven and year eight. The impact of discounting depends on the discount rate being used and the measure life.[7]

2.1.2.4 Differences in Baseline and Energy Efficiency Energy Streams of Benefits

Energy savings calculations are based on the difference between the post energy efficiency state and the assumed baseline. If the baseline equipment has the same level of degradation in performance, then the energy savings factor due to degradation would be 1.0 *and* it would be appropriate to assume constant energy savings over the life of the energy efficiency measure.[8] In fact, if the relative persistence of savings is higher for the energy efficiency measures compared to a baseline consisting of standard measures, then energy savings not only persists, but can increase over time.

[7] For example, if a discount rate of 5% is used, the savings will be reduced by 0.78 multiplied by the energy savings at five years. At 10 years and a 5% discount rate, the new value would be 0.61 multiplied by the energy savings. At a discount rate of 7% for a 10-year period, the value would be 0.51 multiplied by the energy savings.

[8] The report from Peterson et al. (Peterson et al. 1999) is a good example of degradation being measured for both an efficient appliance offered by an energy efficiency program and standard equipment. This study showed that the high-efficiency coils start with and maintain a higher efficiency than standard efficiency coils. The slower degradation rate increases the life of the equipment, and the equipment uses less energy over its operational lifetime. Even though both high-efficiency units and standard units showed performance degradation over time, the lower rate of degradation in the high-efficiency units resulted in a recommended degradation factor exceeding 1.0 in most years. This factor increased from 1.0 to 1.08 over the 20-year expected life of the unit, indicating that savings not only persisted, but actually increased relative to the baseline over the assumed life of the equipment.

These four factors are meant to address the following questions:

- If a persistence study is conducted, is there a reasonable likelihood that the new trend in energy savings over time would be substantively different from the assumptions used in the initial benefit/cost analyses?
- Would the NPV benefits of the program change with a new persistence factor, the discount rate being used, and the likely change in the baseline energy use level that may also be due to performance issues of the baseline equipment?

There may be good reasons to assess persistence, as many factors can influence the stream of energy savings over a three- to 10-year period. The most common of these factors are listed in Table 1.

Table 1. Factors Influencing Persistence

Residential Sector Programs and Measures	Commercial and Industrial Sector Programs and Measures
1. Changes in ownership	1. Business turnover
2. Maintenance practices	2. Remodeling
3. Changes in equipment use	3. Varying maintenance
4. Behavioral changes	4. Operating hours and conditions
5. Occupancy changes	5. Inappropriate installation of equipment
6. Inappropriate installation of equipment	6. Manufacturer performance estimates that do not reflect in-field operating conditions.
7. Manufacturer performance estimates that do not reflect in-field operating conditions.	

Sensitivity analyses using the benefit/cost models can highlight those measures for which adjustments in persistence will have the largest impact. This information can then be used to prioritize persistence evaluation efforts. Thus, before deciding whether additional analyses are needed, test the sensitivity of NPV benefits to potential changes in the persistence of savings. This can help determine whether the impact may be large enough to merit a substantial study effort, or sufficiently small, requiring only a modest retention study.

2.2 State of the Practice in Assessing Persistence

Professional judgment plays a significant role in selecting a method for assessing persistence. The *California Energy Efficiency Evaluation Protocols* (CPUC 2006) has several types of retention, degradation, and measure life/effective useful life (EUL) studies from which to select, based on the priority given to the issue by regulatory staff or other stakeholders.

Evaluators seem to rely on the following two processes for developing estimates of persistence:

- ***Database or Benchmarking Approach.*** This entails developing and regularly updating[9] a database of information on measure life and performance degradation.
- ***Periodic In-Field Studies.*** This entails performing selected in-field studies of program participants from earlier years.

These two approaches are not necessarily mutually exclusive. The database/benchmarking approach is often used when (1) there are a large number of energy efficiency measures, (2) there are concerns about the sample sizes required for in-field studies, and (3) the cost of conducting in-field persistence studies is an issue. Periodic studies may be used for updating a database of measure life and performance degradation. Such studies are also useful when focusing only on those measures that account for a large fraction of the savings. Additionally, in-field studies of program participants that are conducted a number of years after participation provide direct information on persistence of savings for that program.

2.3 Database/Benchmarking Approaches

The three examples of database/benchmarking approaches presented below are based on:

- Engineering judgment
- Experience with the energy efficiency measures
- Information on local and regional conditions to develop tables of measure lives for use in energy efficiency program planning.

These values are often used as deemed values for persistence and applied to produce estimates of the energy savings over time (as inputs to benefit/cost calculations). An assessment of this approach follows the examples. (References to each study are provided for those wanting more information on the methods used beyond the short descriptions provided below.)

2.3.1 Example Study 1: GDS Associates (GDS Associates 2007)

Objective: The measure life values presented in this report were developed to meet the following conditions:

- Accurately reflect conditions for measures installed by energy efficiency programs in the New England states that have supported this research effort
- Satisfy any Independent System Operator-New England (ISO-NE) requirements (for example, for definition and documentation sources)
- Work as common values, accepted by all New England states for the forward capacity market (FCM) (that is, the ISO-NE forward capacity market).

[9] As it is important that these benchmarking studies be updated on a regular basis, the cost of these updates should be included in the cost estimate for using this approach. While these studies may not appear costly on a one-time basis, the effort required to update the database regularly can be significant. This is important, as these databases are sometimes the source of deemed values for measure life and persistence of savings used in evaluation efforts.

Methodology: "Reviewed all secondary data collected and developed a preliminary list of potentially applicable residential and C&I [commercial and industrial] measures. This list was then distributed to program administrator staff within the SPWG [State Program Working Group] for review and to obtain additional program-specific measure life values and associated documentation sources. GDS compiled all responses and developed initial measure life recommendations for SPWG member consideration."

2.3.2 Example Study 2: KEMA (KEMA 2009)

Objective and Methodology: "The principal objective of this study was to update the current measure life estimates used by the Focus Evaluation Team and the Focus Program. **The evaluation team's approach to this study consisted entirely of secondary research**; the team did not conduct primary research, fieldwork, or produce a savings persistence study." (Emphasis added.)

2.3.3 Example Study 3: Energy and Resource Solutions (ERS 2005)

Objective: "The primary goals of the Common Measure Life Study were as follows:

- Define measure life and related terms, such as persistence
- Review the provided table of current measure lives
- Survey other utility energy efficiency programs
- Develop a table of technological measure lives
- Recommend common measure lives and persistence assumptions to be used by the sponsors."

Methodology: "ERS [Energy and Resource Solutions] reviewed the tables of agreed-upon and disputed measure lives provided by the sponsoring utilities. As tasked in our proposal, we researched several sources to use in support of selecting individual measure lives. We first thoroughly researched the CALMAC [California Measurement Advisory Council] database. The CALMAC database provides a public depository for all persistence, technical degradation factor (TDF) and other related studies performed in the State of California. Next, we surveyed many electric utilities and state utility commissions throughout the nation, obtaining other utilities' tables of measure lives. We obtained measure life tables used in 8 states by at least 14 different utilities. Finally, we performed a literature search, referenced technical sources and consulted equipment manufacturers to establish a table of technical lives for each measure. In conjunction with these efforts, we specifically researched the effect of New Construction versus Retrofit status on measure lives, as well as the effect of Small versus Large businesses."

2.4 The Challenges of New Technologies and Measures

The methods in the three examples above have produced useful estimates for a wide number of measures where practical information exists from measure installations and fieldwork. However, new technologies and measures installed less frequently pose greater challenges for this judgment-based benchmarking approach. For many widely implemented energy efficiency measures, both the evaluation work and additional on-site engineering work (such as installation and maintenance) provide a basis for the use of informed engineering judgment. A series of retention/survival rate studies in California—conducted from 1994 to 2006—found that most *claimed* estimates could not be rejected by the in-field studies. However, the in-field studies often had small sample sizes for certain measures and short time frames that did not allow for many failures to occur in the dataset.

Some important measures in these engineering and expert-developed measure life tables may not have fared well. Both residential lighting and commercial lighting have provided a large fraction of savings, and the persistence of these savings has been controversial. Nexus (2008) found that the life for certain lighting measures depends not only on the equipment, but also on the program design.

Skumatz (Skumatz et al. 2009) (Skumatz 2012) critiques the database/benchmarking approach, which is based on engineering judgment combined with literature reviews. Skumatz (2012) identifies strengths and weaknesses in this approach compared to on-site data collection, and she offers suggestions for improving current estimates. Skumatz notes that measure life values existing in tables often vary by more than 25%, and that this has "precisely the same impact on a measure's or program cost-benefit ratio" as savings values that are off by 25%.

While this comment has merit, the measure life and persistence factors will start at 1.0 in the initial years of the program and then gradually change. This change in savings is offset to some degree by the discounting of benefits from five, 10, and 15 years out. Also, this single measure with varying measure life values across engineering-based tables may not represent the composite effective life of a group of measures that make up a program.

2.5 In-Field Persistence Studies (Survey and On-Site Data Approaches)

Methods that make use of in-field data collected on program participants at some point after they participated in an energy efficiency program generally rely on:
Surveys or on-site visits to determine whether the measure is still in place and operable, or, if the measure was removed, when and why[10]

- Statistical analyses using regression-based methods to estimate retention/survival models that produce estimates of the survival or failure rates of energy efficiency measures.

The *California Energy Efficiency Evaluation Protocols*[11] specified these three categories of methods used for in-field studies of persistence:

[10] One reviewer suggested that the surveys referred to in this section should specifically include online approaches. The topics of using online surveys to obtain customer-specific information and combining online surveys with other methods are discussed in the "Survey Research" chapter.
[11] The methodology language from the *California Energy Efficiency Evaluation Protocols* (California Public Utilities Commission 2006) has been adapted to fit the measure life definition and persistence structure used in this chapter. One difference is the use of *persistence* as the overarching term for all types of changes in energy savings over time, which the California Protocols document addresses in the "Effective Useful Life Protocol" section (p. 105). The California *Protocols* still contain the most comprehensive discussion of methods for assessing persistence.

- ***Retention Studies*** provide the percentage of the measures that are in place and operable at a point in time. Retention studies identify technology design, define operable conditions, and describe how operable conditions could be measured.
- ***Measure Life/EUL*** estimates the median numbers of years that the measures installed under the program are still in place and operable. This value is calculated by estimating the amount of time until half of the units will no longer be in place and operable.
- ***Performance Degradation*** uses both technical and behavioral components to measure time-related and use-related changes in energy savings relative to a standard efficiency measure or practice. In general, both standard equipment and energy efficiency equipment become less efficient over time, regardless of the equipment measure life. This factor is a ratio reflecting the decrease in savings due to performance degradation from the initial year savings.

2.5.1 Retention and Measure Life Studies

A retention study determines the number of installed and operable measures at a given point in time. A measure life study is an extension of a retention study, where there is adequate data to allow for the development of a statistical model (commonly called a "survival analysis") to estimate failures that might occur after the data are measured.

Information from the retention model provides an estimate of the measures that were installed and operating at a point in time, which allows the evaluator to calibrate the *claimed* savings and produce adjusted *evaluated* estimates of savings over time. The current estimates of persistence are adjusted to account for the new information *and* the stream of savings over the year. These estimates could, for example, be adjusted in year four to be consistent with the retention study. This ratio for year four would then be used to adjust the savings in all subsequent years.

The measure life estimation methods, which are based on survival analysis, provide more information. However, estimating measure life requires a much larger sample—one that contains an adequate number of both installed and missing (that is, uninstalled or replaced) equipment.

The following are two types of retention and measure life methods, which have been used to estimate the survival models that produce estimates of measure life. (Studies using these methods are described later in this section.)

2.5.1.1 In-Place and Operable Status Assessment (Using On-Site Inspections)

The in-place assessment studies are verified through on-site inspections of facilities. Typically, the measure, make, and model number data are collected and compared to participant program records, as applicable. As-built construction documents may also be used to verify selected measures when access is difficult or impossible (such as wall insulation). Spot measurements may be used to supplement visual inspections—such as solar transmission measurements and low e-coating detection instruments—to verify the optical properties of windows and glazing systems.

Correct measure operation is observed and compared to the project's design intent. Often, this observation is a simple test of whether the equipment is running or can be turned on. However, the observation and comparison can extend to changes in application or sector, such that the operational nature of the equipment no longer meets the design intent. For example, working gas-cooking equipment that had been installed in a restaurant but is now installed in the restaurant owner's home is most likely no longer generating the expected energy savings, so it would not be counted as a program-induced operable condition.[12]

2.5.1.2 Non-Site Methods

Typical non-site methods include telephone surveys/interviews, analysis of consumption data, or the use of other data (such as from energy management systems). The goal is to obtain essentially the same data as would be gotten through an on-site verification; however, there is the potential for collecting inaccurate data, due to a number of factors (discussed in Chapter 11: *Sample Design*).

2.5.1.3 Examples of Retention and Measure Life Studies

Two examples of these types of studies were performed by KEMA and by Nexus Market Research.

- KEMA (KEMA 2004) used a telephone survey to gather information on refrigerators at years four and nine as part of a review of an appliance recycling program.
- Nexus Market Research (Nexus Market Research 2008) used on-site verification data to conduct a measure life study of residential lighting measures.

Both studies provide good examples of collecting information for a basic retention study, and they serve as illustrations of the statistics necessary to estimate a survival model (Allison 1995).[13] Each is discussed below.

Example Study 1: KEMA (KEMA 2004). Conducted with program participants from the years 1994 through 1997, this study looked at retained savings over this period.

For each year, the measure life/EUL estimate reflects the following factors:

- The time at which half of the recycled appliances are from participating premises that have added an appliance
- The time at which half of the recycled appliances would have been out of service without the program influence.

[12] In addition to this language, the *California Energy Efficiency Evaluation Protocols* outlines certain sampling criteria that must be met in California. However, these criteria may vary in accordance with the requirements of different jurisdictions.

[13] To assist evaluators, the *California Energy Efficiency Evaluation Protocols* states: "Multiple statistical modeling packages (SAS, Stata, SPSS, R, S+, and others) provide survival analysis programs. There are several commercial and graduate textbooks in biostatistics that are excellent references for classic survival analysis. One of these used as reference for some of the prior EUL studies in California is the SAS statistical package and the reference *Survival Analysis Using the SAS System: A Practical Guide* by Dr. Paul D. Allison, SAS Institute, 1995. Several model functional forms are available and should be considered for testing. These forms include logistic, logistic with duration squared (to fit expected pattern of inflection point slowing of retention losses), log normal, exponential, Weibull, and gamma."

The KEMA study illustrates one way in which the *claimed* and *evaluated* measure life values can be used. As stated in the study:

> For each of the program years from 1994 through 1997, both refrigerators and freezers have a claimed (or *ex ante*) estimate of measure life/EUL of six years, which has been used in the earnings claims to date. A measure's evaluated measure life/EUL is the value estimated by a persistence study. If a measure's claimed measure life/EUL is outside the 80% confidence interval, the measure's evaluated measure life/EUL may be used for future earnings claims. Otherwise, the measures claimed value will continue to be used in earnings claims.

Figure 2 is a replication of Table E-1 from the KEMA study, which shows the comparison between the *claimed* and *evaluated* measure life/EUL estimates. In this case, the measure life results showed that the program was underestimating the measure life/EUL values *and* that the realization rate exceeds 1.0.

Table E-1
1994-1997 Appliance Recycling Program
Summary of Effective Useful Life Estimates

			EUL (years)					
						80% Confidence Interval		
Program Year	Measure	End Use	Ex Ante	Ex Post (estimated from study)	Adopted ex post (to be used in claim) Lower Bound	Lower Bound	Upper Bound	EUL Realization Rate (adopted ex post/ex ante)
1994	Freezer	Refrigeration	6.0	8.0	8.0	8.0	11.0	1.33
	Refrigerator		6.0	8.0	8.0	8.0	11.0	1.33
1995	Freezer	Refrigeration	6.0	8.0	8.0	8.0	11.0	1.33
	Refrigerator		6.0	8.0	8.0	8.0	11.0	1.33
1996	Freezer	Refrigeration	6.0	8.0	8.0	8.0	8.0	1.33
	Refrigerator		6.0	8.0	8.0	8.0	8.0	1.33
1997	Freezer	Refrigeration	6.0	8.0	8.0	8.0	8.0	1.33
	Refrigerator		6.0	8.0	8.0	8.0	8.0	1.33

Figure 2. KEMA (2004) Table E-1

Example Study 2: Nexus Market Research (2008). This study examined the measure life of lighting products distributed through energy efficiency programs in New England.

The definition of measure life is the same as presented above in *Addressing Persistence* and used in the Energy and Research Solutions (2005) example application presented above. Specifically, Nexus states that:

> [T]he measure life estimates do not distinguish between equipment life and measure persistence; our estimates—one for each measure category—include both those products that were installed and operated until failure (that is, equipment life) as well as those that were retired early and permanently removed

from service for any reason, be it early failure, breakage, or the respondent not liking the product (that is, measure persistence).

Nexus drew a random sample of participants based on the type and number of products they had obtained through the programs. The report states, "We collectively refer to these sample products as the 'measure life products.'"

Auditors visited 285 homes to inventory lighting products, and Nexus designed a respondent survey to learn more about the measure life products and other lighting products found in the home. These survival analyses were based on the following methods and, ultimately, Nexus used estimates resulting from Method 3.

- Method 1: Measure Life Tables
- Method 2: Logit Regression
- Method 3: Parametric Regression Models of Survival Analysis.

The results showed that the measure life for compact fluorescents (CFLs) varies by program design (that is, whether the program was coupon-based, direct install, or a markdown at a retail facility). The results of the Nexus (2008) study are shown in Table 2.

Table 2: Nexus (2008) "Recommended Estimates of Measure Life—Decimals"

Product	Measure Life	80% Confidence Interval	
		Low	High
Coupon CFLs	5.48	5.06	5.91
Direct Install CFLs	6.67	5.97	7.36
Markdown CFLs (all states)	6.82	6.15	7.44
Coupon and Direct Install Exterior Fixtures	5.47	5.00	5.93
Markdown Exterior Fixtures	5.88	5.24	6.52
All Interior Fixtures	Continue using current estimates of measure life		

Nexus deemed a representation of the results—at an 80% confidence interval—as being accurate enough for the purposes of this study. Nexus recommended measure life estimates for three measures: one for compact fluorescent lamps (CFLs; coupon, direct install, and markdown)[14] and two for exterior fixtures (markdown and all other programs).

Nexus did not recommend an estimate of measure life for interior fixtures, as the timing was too early in the measure lifecycle to provide a reliable estimate. This occurs with a number of measure life studies that are conducted too early (before there have been enough failures or un-installs to allow for statistical modeling of measure life).

2.5.2 Examples of Degradation Studies

While there are few reports that directly focus on the degradation of savings, two types of studies are available, and they are described below:

[14] Due to the diversity of program types throughout the region, Nexus used the term "markdown" to refer to both markdown programs (offered in all of the states) and buy-down programs (offered in some of the states).

- Focusing on technical degradation (one of the clearest examples is by Proctor Engineering in 1999 [Proctor Engineering 1999])
- Performing billing analyses at some point after participation to capture all of the factors that impacted persistence of savings. (In 2011, Navigant performed a billing analysis of a customer information program, which was used to examine persistence of impacts across two years for a behavioral program. [Navigant 2011])

Example Study 1: Proctor Engineering (Proctor Engineering 1999). The purpose of this project was "to examine the relative technical degradation of demand side management (DSM) measures compared to standard efficiency equipment. This project covers two major DSM measures: commercial direct expansion air conditioners (Comm. [direct expansion] DX AC) and EMS [energy management systems]."

Proctor Engineering's methodology involved establishing a time-series estimate—derived from available research—for condenser and evaporator coil fouling rates. Proctor used laboratory testing to modify the estimated fouling rates and establish a profile for coil fouling. It tested both high-efficiency and standard efficiency coils in a controlled laboratory environment, and both were subjected to continuous fouling. Proctor then monitored the efficiency of the air conditioner at various intervals to document the effects.

This study found that (1) the impact on standard equipment was greater and (2) the high-efficiency units actually had a higher level of savings persistence. The end result was that "testing shows that the TDF [technical degradation factor] for this measure is greater than one." This is an example of degradation needing to be conducted with reference to standard efficiency equipment. Energy efficiency measures may have performance degradation, but so does standard equipment. If the energy efficiency measures have a lower rate of degradation, then savings increase (as measured against the standard equipment baseline).

To assess EMS, Proctor used an on-site methodology rather than laboratory testing. The research data showed that although there is some EMS savings degradation at some locations, other locations show increasing savings. Some of the causes for this persistence are:

- No instances of disconnected or non-operational EMSs were found.
- The vast majority of EMSs appeared to be operated in a competent and professional manner.
- EMS operators had found that the EMS was a useful tool in performance of their jobs.

Proctor Engineering contrasted its work with other EMS studies showing greater degradation due to operational issues. Proctor explained the comparatively high level of persistence it found as being due to the high interest of the program participants in saving energy. The more random group of facilities in the comparison may not have been involved in EMS-related energy efficiency programs.

Proctor also conducted a billing analysis to confirm these findings. For this billing analysis, it combined the consumption data from all of the sites and then estimated the persistence of savings over time. The regression process provided statistically significant estimations at the 95% level.[15]

The primary purpose of this research was to establish the TDFs, estimated for each measure. The results from Proctor's study, seen in Figure 3, shows that the degradation factors are greater than 1.0 for the high-efficiency DX AC equipment. This indicates the degradation was less for the high-efficiency DX AC equipment than for the standard efficiency equipment.

Table ES-1 TDF

Year	EMS	Comm DX AC
1	1.00	1.00
2	1.00	1.00
3	1.00	1.00
4	1.00	1.01
5	1.00	1.01
6	1.00	1.01
7	1.00	1.01
8	1.00	1.01
9	1.00	1.01
10	1.00	1.02
11	1.00	1.02
12	1.00	1.02
13	1.00	1.02
14	1.00	1.02
15	1.00	1.02
16	1.00	1.02
17	1.00	1.02
18	1.00	1.02
19	1.00	1.06
20	1.00	1.08

Figure 3. Proctor Engineering (1999) Table ES-1

Still, the difference is small through year 18, and this size of effect might not show up in benefit/cost analyses due to the discounting required to obtain an NPV of savings benefits.

Example Study 2: Navigant (Navigant 2011). This study examined the short-term persistence of a behavioral information program using billing data across multiple years, as short-term persistence may be an important factor for these programs.

The program was designed to assist and encourage customers to use less energy. These types of programs are increasing in the industry; for example, OPOWER, Inc., offers residential customers regular Home Electricity Reports about their electricity consumption to help those customers manage their electricity. In combination with other information, these reports compare

[15] References to statistically significant results in regression analyses must be carefully interpreted. The analysis may have been a test to determine if the effect was significantly different from zero (±100% precision). Alternatively, the test may have actually established a precision level of ±10% or another level of precision, (for example, 30%). A statement of statistically significant results should be accompanied by an explanation for interpreting that statement in terms of the level of precision being used in the test of significance.

a household’s electricity use to that of its neighbors and then suggest actions to reduce electricity use. It is hypothesized that presenting energy use in this comparative fashion creates a social nudge that induces households to reduce their consumption.

Navigant evaluated the first 29 months of the program, with an emphasis on the second program year. The following main research questions were addressed in the evaluation and presented in this report:

- Does the program continue to generate savings?
- What is the trend in program savings? Is there a ramp-up period to savings? If so, for how long? Are savings now relatively stable, increasing, or falling?
- Do program savings increase with usage?

The evaluation of this program entailed developing a random control group and conducting a fixed-effects regression analysis, which is a common evaluation method. This regression method is discussed in the “Whole House Retrofit” chapter of this UMP report.

Navigant’s results showed that the effects of slightly more than 2% of the energy savings persisted across the 29 months examined in the study, after an initial ramp-up period of approximately 10 to 12 months. The small effect size required a large sample of customers for the regression analysis to produce reliable results. For this behavioral program evaluation, there were more than 20,000 treatment customers and a control group of more than 30,000 customers. Thus, large samples are needed to identify small effect sizes from energy efficiency programs.

This regression framework can be applied to a third and fourth year of data to assess longer-term participation.

2.6 Persistence Recommendations and Conclusions

Evaluators address the issue of persistence of savings from energy efficiency programs because of the impact that the stream of savings estimates has on the benefit/cost tests of measures and programs. While some measure life values are estimated at more than 20 years, most benefit/cost assessments are estimated out at least 10 years or, more commonly, 15 to 20 years.

The approaches discussed in this chapter include methods to address measure life and savings performance, which may be impacted by operating conditions, behavioral changes, turnover in building occupancy, changes in measure use, and other factors. To date, the tools and methods that make up the recommended tool kit for evaluators include:

- Benchmarking and database development for measure life values and savings persistence
- On-site analyses of equipment
- Survey methods for select measures amenable to survey techniques
- Single-year estimations of equipment retention and operation
- Multiyear statistical analyses based on survival models

- Technical degradation studies based on engineering review
- Technical degradation based on laboratory testing
- Billing analyses that capture overall persistence (that is, that assess savings directly and capture all changes in savings for the time period being analyzed).

The review of methods illustrates the different ways persistence can be addressed. Research is continuing in this area, and methods have been adopted in different jurisdictions. As with any area of evaluation, there will always be improvements. The *Appendix* to this chapter presents tables outlining program and measure persistence study challenges and issues.

The balance of this section presents practical recommendations for assessing the persistence of savings. The goal of evaluation is to help stakeholders make good decisions about investments in energy efficiency programs, and this requires both an understanding of the techniques and applied judgment.

2.6.1 Recommendations

1. ***Before determining whether to undertake a large-scale persistence study of a program or measure (or even to undertake such a study at all), consider whether the results of the study are likely to have a material impact on the economics of the program.*** Persistence of savings refers to the stream of savings expected from a measure or program over a period of years. If the study's revised persistence of savings is expected to be small and to occur 10 or more years or more in the future, then the impact of that change may not have a large effect on the cost-benefit economics.

   Keep these considerations in mind when deciding:

   - Benefit-cost tests are based on NPVs that discount the streams of benefits and costs. A change in measure life by a year or two *and* changes for long-lived measures may not have much impact after they are discounted.
   - The performance degradation of energy efficiency measures should be assessed relative to that of the standard efficiency equipment, as both will have performance degradation. The difference between these two values determines the impact on savings.

2. ***Select the methodology that best fits the individual circumstances of the measure/program being evaluated.***

   - Pick the method most appropriate to the magnitude of the effect expected. Before conducting the study, take a forward-looking view of what might be learned. While this may seem difficult, researchers across the evaluation community and the industry make these decisions on a regular basis. The key is to ensure that the information produced is worth the effort expended to produce it. The goal is to obtain information that decision makers need for making good decisions regarding energy efficiency investments.

- Measures that may have persistence impacts within the first three to seven years are the most important to study because of their near-term effects and their potential to influence the benefit/cost tests and program designs.
- As benchmarking uses the expertise of engineers who have been working in the field for years, it may be a good approach for many measures, particularly given the large number of measures across all energy efficiency programs. However, past work can be improved upon through the use of more systemized approaches, such as a Delphi-type of analysis.[16]
- Although the billing analyses method addresses the issue of persistence most comprehensively, there are cautions to consider. The effect may be small, which will require large sample sizes. Also, it may be difficult to control for other factors outside the program that cause changes in energy use across a five- or 10-year period. Where quality data exist, a billing analysis is a good method for assessing persistence, but it requires an appropriate data platform for it to be reliable.[17]

3. ***It is important to be open to the new methods and approaches being developed.*** Specifically, a panel of participants established at the time of program participation could be used in cross-sectional, time-series models. This involves incorporating the evaluation of persistence in program design and implementation planning. This type of forward thinking will make persistence easier to address, particularly in near-term years when it is most important.[18]
4. ***Certain types of persistence studies, particularly database/benchmarking approaches, might best be addressed on a regional basis that includes numerous specific programs.*** Assessing persistence across a number of regional programs can provide information on the influence of program design on persistence, which might not be found using a series of program-specific studies. In identifying these regional opportunities, it is important to consider the influence of program design on persistence. (For example, in the study Nexus performed across New England in 2008, program-specific elements had a large influence on the persistence of lighting measures.)

[16] Skumatz (2012) presents a number of ways these studies can be improved, including the use of Delphi approaches. An expert-panel approach was used in an evaluation of the Northwest Energy Efficiency Alliance's market transformation programs by Violette and Cooney (Violette and Cooney 2003).

[17] Billing data analyses that try to estimate small effects reliably (for example, 2% savings) without the required sample sizes and accurate data for the independent variables (that is, little measurement error) have often not been successful. Quantum (Quantum 1998) discusses this issue in the context of using a billing analysis to assess persistence for new home construction.

[18] Panel data methods are suggested as a potential approach in both Skumatz (Skumatz 2012) and Nexus (Nexus 2008).

3 Other Evaluation Issues

This section briefly addresses these evaluation issues: (1) synergy; (2) errors in variables, measurement error, and program tracking; (3) dual baselines; and (4) rebound.

3.1 Addressing Synergies Across Programs

Evaluators are often asked about potential synergies across programs. For example, certain information programs may result in direct savings impacts, but the programs may also be designed to lead participants into other programs. In addition, there may be effects across programs. For example, a whole-house retrofit program may influence the uptake of measures offered in other residential programs. These synergies are useful for designing programs and portfolios. Synergies that increase the overall savings from a portfolio of programs are valuable even if one specific program has lower savings due to these synergies.

The industry practice is to use approximate information to assess the relative importance of synergies. Even this level of analysis has generally been limited in evaluations. However, useful information on synergies can be developed by having evaluators:

1. Identify what they believe may be positive and negative synergies (that is, direction)
2. Determine the rough magnitude of these synergies by benchmarking them as a fraction of the programs' savings.

With this material, portfolio models designed to assess the importance of synergies can produce information useful for assessing investments in energy efficiency and future program/portfolio designs.[19]

[19] This approach does not have to be information intensive in terms of developing useful data for analyzing synergies and benchmarking their magnitude. Two pieces of information are needed: (1) an estimated range of effects, for example, from 5% of program savings to 20% of program savings; and (2) an estimate of where the most likely value falls within this range. Based on these three points—the lower bound, the upper bound, and an estimate of where within this range the most likely value falls—Monte Carlo methods can be used to test the importance and sensitivity of program impacts to identified synergies using Excel-based tools. An example of this range-based method can be found in Violette and Cooney (Violette and Cooney 2003), and a version of this method is discussed in EPRI (EPRI 2010, p. 5-4). This information can be used by the program administrator to inform the design of future energy efficiency portfolios.

3.1.1 Conclusion

At the present time, the state-of-the practice involves identifying and assessing the potential importance of specific synergies across programs, although this is not always requested of evaluators. If assessing synergies becomes part of an evaluator's reporting requirements, the evaluator could modify surveys to provide useful information on potentially important energy efficiency program design considerations.[20]

3.2 Errors in Variables, Measurement Errors, and Tracking Systems

This section outlines the issues of errors in the input variables to an energy savings calculation. Such errors could be caused by an incorrect engineering calculation or by inaccurate values of the independent variables used in the regression analyses.

It is important that evaluators consider the accuracy of the input data and use the best quality data possible. In this context, data accuracy issues include data that are unbiased on average, but are subject to measurement error. Biased data clearly poses issues for any analysis; however, measurement error in itself poses challenges for evaluation. This is true even when the measurement error may be uncorrelated with the magnitude of the value of the variable, and the error may be equally distributed above and below the true value.

Program implementers need to be aware that the designs of the data tracking system and the data collection processes have a substantial influence on the accuracy and reliability of data. In turn, the accuracy and completeness of the data influence the estimated realization rates and the ability to achieve the target levels of confidence in these estimates.

While errors in variables can bias the evaluation results either up or down, there are several practical factors in energy efficiency evaluations that tend to result in lower realization rates and lower savings estimates. A typical realization rate study uses information from the tracking system to verify that the equipment is in place, working as expected, and achieving the energy savings predicted in the tracking system. Tracking system errors can include not properly recording the site location, contact information, equipment information, location where the equipment is installed, and the operating conditions of the equipment. This will make any associated field verification more difficult and the variance around the realization rate greater.

Different data issues will have different impacts on the estimates; however, improved data quality will usually decrease the variance of the realization rate estimate and increase confidence and precision. When stakeholders have set high target confidence-and-precision levels, it is important to track accurately the essential data (such as the installed measures' location, size, model number, date, contact person) required to produce the initial tracking system estimate of savings at that site.

[20] One reviewer of this chapter pointed out the potential complexities of determining program-specific synergies and their direction "...to the extent that synergies are increasingly observed or acknowledged, policies regarding the use of individual program cost-benefit analysis results for justifying the retention of programs may need to be changed in favor of portfolio level benefit cost analyses." This section was not intended to delve into benefit-cost methods. However, increased attention on synergies across programs is likely to prove useful. Monte-Carlo models that use different scenarios regarding the magnitude and direction of synergies can help assess the robustness of program and portfolio cost-effectiveness.

The issue of errors in variables and measurement error can be important.

- Kennedy (2003) states: “Many economists feel that the greatest drawback to econometrics is the fact that the data with which econometricians work with are so poor.”
- Similarly, Chen et al. (Chen et al. 2007) states: “The problem of measurement errors is one of the most fundamental problems in empirical economics. The presence of measurement errors causes biased and inconsistent parameter estimates and leads to erroneous conclusions to various degrees in economic analysis.”

Errors in measuring the dependent variable of a regression equation are incorporated in the equation’s error term and are not a problem. The issue is with errors in measuring the independent variables used in a regression model. This violates the fixed independent variables assumption of classical linear regression models: the independent variable is now a stochastic variable.[21] A good source for approaches to address the errors-in-variables issue is Chapter 9 in Kennedy (2003).

The program tracking system data used in regression analyses can be a source of potential issues. For example, the inability to track customer participation in multiple programs can cause a number of problems. In these instances, data can be very accurate at the program level, but there is no mechanism to ascertain the effects of participating in multiple programs. For example, if a billing analysis is being conducted of a high-efficiency residential heating, ventilating, and air-conditioning (HVAC) replacement program but the tracking system is not linked to the residential audit and weatherization program that feeds participants into the HVAC program, this will cause bias. When customers first participate in a feeder program but that information is not conveyed in the tracking system used by the HVAC evaluator, then the HVAC program’s savings analysis will be biased, most likely on the low side.

Another well-known errors-in-variables issue relates to models that use aggregate data on DSM expenditures and energy consumption in analyzing the relationship between expenditures on energy efficiency activities and changes in energy use.[22] Developing the appropriate datasets poses challenges. For example, Rivers and Jaccard (2011) note that:

> [O]ur data on demand side management expenditures include all demand side management—in particular it includes both load management expenditures as well as energy efficiency expenditures. Since load management expenditures are not aimed at curtailing electricity demand explicitly… (p. 113).

The report then states that they do not believe this is a problem since

> ...utilities that were able to provide us with data (as well as in US utilities), load management expenditures amounted to less than 25% of the total, so error in our

[21] The assumption is that observations of the independent variable can be considered fixed in repeated samples (that is, that it is possible to repeat the sample with the same independent variable values; [Kennedy 2003, p. 49]).
[22] Two recent publications with examples of this are Rivers and Jaccard (Rivers and Jaccard 2011) and Arimura et al. (Arimura et al. 2011).

> estimates should not be too severe, and in particular should not change the nature of our conclusions.

The authors may be correct, but their assessment was based on judgment with little real analysis of the degree of the issue.

The work by Rivers and Jaccard (Rivers and Jaccard 2011) and by Arimura et al. (Arimura et al. 2011) illustrates the degree of effort often required to develop a useful set of aggregate state/province-level data or utility-level DSM. Using the Energy Information Administration forms, Arimura states: "The original data set has many observations with missing values for DSM spending, even after our meticulous efforts to find them from various sources."[23]

Another issue concerns the fact that numerous states have both utility and third-party program providers, which complicates the development of data that can be used to examine the relationship between utility energy efficiency program expenditures and aggregate energy consumption.

Attenuation bias is a potential issue when there is measurement error in the independent variables used in regression analyses. Simply stated, the implications are these: (1) more noise in the data due to measurement errors will make it more difficult to find significant impacts and (2) those impacts will tend to be biased downwards.[24]

Attenuation bias can be a problem in regression models using independent variables that might have large numbers of measurement errors due to:

- Differences in reporting of values in databases compiled across utilities
- Assignment/allocation of values at a utility service territory level down to a county level to create more observations.

Chen et al. (Chen et. al 2007, 2011) and Satorra (Satorra 2008) present a graphical example of this bias using a measurement error model developed for a simple one-variable regression.

- Using the model Y = βX + *e* and
- having X measured with error,
- the measurement error model X = x + u, with x uncorrelated with u, var(X) = var(x) + var(u) can be used to assess the reliability of the estimated coefficient.

The reliability of X is defined as rel = 1 - var(u)/var(X) (which results in a number between 0 and 1).

[23] See footnotes 15, 16 and 17 in Arimura et al. (2011) for a discussion of the challenges they addressed in developing values of the key variables (that is, the utility's energy efficiency expenditures that could explain changes in energy use and be used to assess cost-effectiveness in terms of cost per kWh saved).
24 This is not a new problem. Chen (2007 and 2011, p. 901) discusses how one of the most famous studies in economics had to address attenuation bias. In his famous book A Theory of the Consumption Function, Milton Friedman (Friedman1957) shows that, because of the attenuation bias, the estimated influence of income on consumption would be underestimated.

Satorra performed a set of simulations for a sample size equal to 10 and used different values for the reliability of the regressor X: 1 (accurate), 0.86, 0.61, and 0.50 (considerable measurement error).

Each simulation is shown in Figure 4.

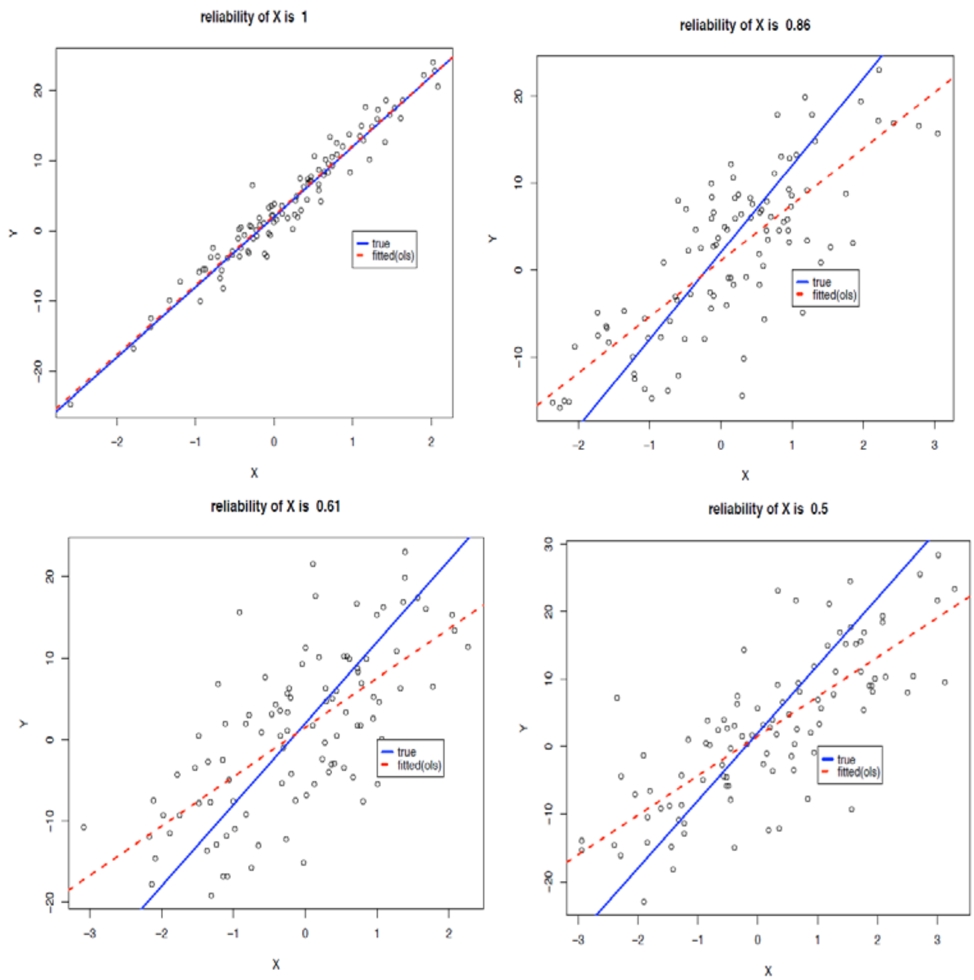


Figure 4. Satorra (2008) Simulation Results

As shown in Figure 4, the bias in the coefficient increases as the reliability of X decreases (that is, measurement error increases), even if this measurement error is uncorrelated with the variance of X. The slope of the coefficient declines as the reliability of X declines. This represents the attenuation bias associated with measurement error.

3.2.1 Conclusion

Issues associated with measurement error are often unavoidable in applied regression analysis. On occasion, data collected for one purpose with one level of accuracy may be used as a variable in a model testing for different types of effects. The solution is to reduce measurement error in the independent variables (the regressors) as much as possible.

Errors in variables, measurement errors, and general issues with data in tracking systems will make it more difficult for the evaluator to identify energy savings at a desired level of confidence. Kennedy (2003) states, "In the spirit of fragility analysis, econometricians should report a range of estimates corresponding to a range of values of measurement variance." Kennedy presents examples of how this can be accomplished, but this extra effort is best reserved for large-scale efforts, and it goes beyond current industry standard practice in energy efficiency evaluation.

Nevertheless, having a good data platform from which energy efficiency savings are evaluated is important and needs more emphasis in practical evaluation work.

3.3 Dual Baselines

There are several evaluation issues caused by changes—during the lifetime of that measure—in the baseline against which savings are estimated. One issue, remaining useful life (RUL), occurs when a program is focused on replacing existing (lower-efficiency) equipment with energy efficiency equipment before the old equipment ceases to function *or* before it would otherwise have been replaced. The savings could be:

- Calculated simply as the difference between energy use for the replaced measure and the new energy efficiency measure or
- Based on the difference between the new standard measures available in the market as compared to the new energy efficiency measure.

These savings would be constant for the assumed life of the measure—that is, no adjusted baseline for that measure is considered for the period after the RUL.

In theory, the use of two baselines can be argued to be the appropriate approach in certain applications. The baseline for the replaced low-efficiency measures that still had useful life would be the difference in efficiency between the replaced measure and the high-efficiency measure for the RUL of the replaced measure. For the period *after* the replaced measure's RUL, the baseline should shift to the difference between the installed high-efficiency equipment and the currently available standard equipment. (This would be the baseline for the balance of the assumed life of the new high-efficiency measure.) In practice, this is not often done. (See the conclusions for this section).

A similar situation occurs when a replacement is made of equipment that has a measure life spanning a point when a new code requires higher-efficiency equipment. In this case, evaluators must decide whether the baseline should be the efficiency of the equipment replaced and, in that event, change to a new baseline after the new code or standard is adopted. In general, the working assumption is that the baseline should reflect the energy use of the replaced equipment. If, however, that equipment would have been replaced within a few years by new equipment that meets the new code, then there is a question about whether the baseline should shift.

3.3.1 Conclusions

These dual baseline questions are beginning to receive more attention. Two opinions are expressed in the literature:

- The first and most common is that the complexities and uncertainties entailed in estimating the RULs of the equipment being replaced are excessive compared to their effects on energy savings calculations.
- The second opinion is that dual baseline the issues are important to address for some certain select measures, such as lighting, where the impacts may be large.

These dual-baseline issues have been addressed in some program evaluations, but have not generally been viewed as important for overall energy efficiency program evaluation because of their complexity and uncertainty regarding customer actions. However, the topic of dual baselines deserves more research to assess those specific situations in which accounting for the two baselines might have a substantive effect on energy savings.

3.4 Rebound Effects

Rebound occurs when the costs of using energy are reduced due to energy efficiency programs. When families spend less money to cool their home in the summer because of more efficient equipment, they might change their temperature set point to increase their comfort and their energy use.

Rebound is discussed in the literature according to the following two types:

- ***Type 1: Rebound is used essentially synonymous with take-back*** and happens at the participant level. It involves the question of whether participants who experience lower costs for energy because of an energy efficiency program measure—such as the installation of a high-efficiency air conditioner—then "take back" some of those savings by using more energy.[25]
- ***Type 2: Rebound takes place in the larger economy*** because energy efficiency programs have reduced the cost of energy across a number of uses, stimulating the development and use of energy-using equipment.

With the exception of low-income programs, Type 1 rebound has not been found to be significant in most energy efficiency program evaluations.[26] When consumers match marginal benefits with marginal costs, the concepts of bounded rationality and compartmentalized decision making are being recognized as one theory of consumer behavior and decision

[25] A reviewer pointed out that, for many customers, the lower costs of energy are not reflected in the price of a kWh or a therm of natural gas. Instead, customers use less energy, resulting in a lowering of their monthly bills. This results in customers spending less on energy over the course of a season or year.

[26] This chapter is focused on energy efficiency programs. Take-back is more common in demand response and load management programs where AC units or other equipment are cycled to reduce peak demand for several hours on a few select days. This can result in a warming of the house or building, and the equipment automatically runs a bit more after the cycling event to return the temperature to the original set point. More efficient operational and cycling designs for AC load management programs can greatly reduce take-back, and take-back is a more common effect for event-based load management programs than for energy efficiency programs that influence all hours of a season.

making.[27] (This is contrary to pure economic theory.) Consumers optimize, but only to the point when the complexity of the decision and the cost of the information become too high. For example, although the efficiency of an air-conditioning (AC) unit varies daily with temperature and load; however, a consumer setting the thermostat on the AC unit is probably not going to examine the cost of running that unit each day and then adjust the thermostat accordingly.

Most customers set their thermostats at a comfortable level, regardless of whether they participate in an AC equipment program (whether for maintenance or new equipment) that increases the energy efficiency of the unit. In other words, consumers generally do not change their thermostat setting as a result of participating in an energy efficiency program.

Low-income customers can be the exception, as they may change their thermostat set points for both AC and heating after participating in an energy efficiency program designed to increase the efficiency of the equipment. The change in energy price is more important to low-income customers, who may have been sacrificing comfort to meet their household budget before they participated in the energy efficiency program. Lowering the costs of AC and heating may allow them to set their thermostats at a level that provides more comfort, which may result in greater energy use for this participant segment. While this may cause an increase in the overall energy use for these low-income customers, it can provide a large welfare gain and even improved health and safety for low-income customers.

Going beyond the program participants' actions, Type 2 rebound assesses the economy as a whole, as lowering the cost of energy through aggressive energy efficiency programs may make energy more economical for many new uses. There has been a recent resurgence of interest in this type of rebound, but a full analysis is beyond the scope of this chapter, which focuses on energy efficiency program evaluation. (Gavankar and Geyer [2010] present a review of this larger rebound issue.) There is substantial literature on this economy-wide concept of rebound, and addressing most of the key theses in the discussion requires economy-wide models with energy as one of the inputs for the a wide variety of products and services.[28]

Searching on the terms "energy efficiency" and "rebound" results in many policy papers that present theses on how rebound may be an influence in the larger economy. The issue seems not to be economic welfare, but other policy goals. Using resources as efficiently and cost-effectively as possible always seems like a good policy, unless there is some other constraint. Reducing the cost of energy and allowing people to use energy in additional applications may increase overall welfare. Still, if the goal is to not increase energy use at all, then the downside of reducing energy costs may be concerns about carbon emissions. (It is not the purpose of this chapter, however, to detail this literature, other than noting it exists and offering some practical places to begin a review.)

[27] The primary reference for this concept is Simon (Simon 1957), but it is also discussed in Kahneman (Kahneman 2003).

[28] Other references to discussions of the rebound effect can be found in Vaughn (2012) and in Burns and Potts (2011). Other references are Tierney J. (2011), which presents the issue of rebound as being important, and a counterpoint paper by Afsah (2011.

Using resources as efficiently as possible should be a good start towards any policy designed to reduce energy consumption that may contribute to carbon emissions. This policy could complement pricing and other policies designed to reduce energy use. Starting from a platform of efficient energy use should not hinder the applicability of other policies.

4 References

Allison, P.A. (1995). Survival Analysis Using the SAS System: A Practical Guide. SAS Institute.

Arimura, T.H.; Li, S.; Newell, R.G.; Palmer, K. (2011). Cost-Effectiveness of Electricity Energy Efficiency Programs. National Bureau of Economic Research. NBER Working Paper No. 17556. www.nber.org/papers/w17556.

Afsah, S, K. Salcito and C.Wielga (2011), "Energy Efficiency is for Real, Energy Rebound a Distraction," CO2 Scorecard Research Notes, January. http://co2scorecard.org/Content/uploads/Energy_Efficiency_is_for_Real_CO2_Scorecard_Research_Jan_11_12.pdf

California Public Utilities Commission (CPUC). (2006). California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals. Oregon: TecMarket Works. www.calmac.org/publications/EvaluatorsProtocols_Final_AdoptedviaRuling_06-19-2006.pdf.

Chen, X.; Hong, H.; Nekipelov, D. (2007). "Measurement Error Models." Prepared for the Journal of Economic Literature. www.stanford.edu/~doubleh/eco273B/survey-jan27chenhandenis-07.pdf.

Chen, X.; Hong, H.; Nekipelov, D. (2011). "Nonlinear Models of Measurement Errors." Journal of Economic Literature. (49:4); pp. 901–937. http://cowles.econ.yale.edu/P/cp/p13a/p1344.pdf.

Electric Power Research Institute (EPRI) (January 2010), Methodological Approach for Estimating the Benefits and Costs of Smart Grid Demonstration Projects. Palo Alto, CA: 2010. 1020342. www.smartgridnews.com/artman/uploads/1/1020342EstimateBCSmartGridDemo2010_1_.pdf.

Energy and Resource Solutions (ERS). (November 17, 2005). Measure Life Study. Prepared for The Massachusetts Joint Utilities.

Friedman, M. (1957). A Theory of the Consumption Function. Princeton University Press. www.nwcouncil.org/energy/rtf/subcommittees/comlighting/Measure%20Life%20Study_MA%20Joint%20Utilities_2005_ERS-1.pdf.

GDS Associates (2007). Measure Life Report: Residential and Commercial/Industrial Lighting and HVAC Measures. Prepared for The New England State Program Working Group for use as an Energy Efficiency Measures/Programs Reference Document for the ISO Forward Capacity Market. http://neep.org/uploads/EMV%20Forum/EMV%20Studies/measure_life_GDS%5B1%5D.pdf.

Gavankar, S.; Geyer, R. (June 2010). The Rebound Effect: State of the Debate and Implications for Energy Efficiency Research. Institute of Energy Efficiency (UCSB). http://iee.ucsb.edu/files/pdf/Rebound%20Report%20for%20IEE-UCSB_0.pdf.

Kahneman, D. (2003). “Maps of Bounded Rationality: Psychology for Behavioral Economics.” *The American Economic Review.* (93:5); pp. 1449–75. www.econ.tuwien.ac.at/lotto/papers/Kahneman2.pdf.

KEMA. (2004). Residential Refrigerator Recycling Ninth Year Retention Study. Study ID Nos. 546B, 563, Madison. www.calmac.org/results.asp?t=2.

KEMA. (August 25, 2009). Focus on Energy Evaluation: Business Programs: Measure Life Study Final Report. Prepared for the Public Service Commission of Wisconsin. www.focusonenergy.com/files/Document_Management_System/Evaluation/bpmeasurelifestudyfinal_evaluationreport.pdf.

Kennedy, P. (2003). A Guide to Econometrics. Chapter 9. MIT Press.Navigant Consulting, Inc. (February 20, 2011). Evaluation Report: OPower SMUD Pilot Year 2. Presented to OPower. http://opower.com/uploads/library/file/6/opower_smud_yr2_eval_report_-_final-1.pdf.

Nexus Market Research, Inc. (June 4, 2008). Residential Lighting Measure Life Study. Cambridge: New England Residential Lighting Program. www.puc.nh.gov/Electric/Monitoring%20and%20Evaluation%20Reports/National%20Grid/121_NMR_Res%20ltg%20measure%20life.pdf.

Peterson, G.; deKieffer, R.; Proctor, J.; Downey, T. (1999). Persistence #3A: An Assessment of Technical Degradation Factors: Commercial Air Conditioners and Energy Management Systems, Final Report. CADMAC Report # 2028P.

Burns, C; Potts, M (2011). The “Rebound Effect”: A Perennial Controversy Rises Again. Solutions Journal: Spring 2011 (Vol. 4, No. 2) http://www.rmi.org/TheReboundEffectAPerennialControversyRisesAgain.

Vaugh, K. (2012). Jevons Paradox: The Debate That Just Won't Die – Response by Amory Lovins, RMI Outlet, March 20. http://blog.rmi.org/blog_Jevons_Paradox

Proctor Engineering. (1999). Summary Report of Persistence Studies: Assessments of Technical Degradation Factors, Final Report. CADMAC Report #2030P. www.calmac.org/publications/19990223CAD0003MR%2EPDF.

Quantum Consulting. (1998). PG&E Statewide Multi-Year Billing Analysis Study: Commercial Lighting Technologies Final Report. Berkeley.

Rivers, N.; Jaccard, M. (2011). “Electric Utility Demand Side Management in Canada.” The Energy Journal. (32:4); pp. 93-116. http://bwl.univie.ac.at/fileadmin/user_upload/lehrstuhl_ind_en_uw/lehre/ws1112/SE_Int._Energy_Mgmt_1/DSMCanada.pdf.

Satorra, A. (2008). Theory and Practice of Structural Equation Modeling, Department d'Economiai Empresa. Barcelona. http://statmath.wu.ac.at/courses/TPStrucEqMod/errorinvariables.pdf.

Simon, H. (1957). A Behavioral Model of Rational Choice in Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting. New York: Wiley.

Skumatz, L.A. et al. (2009). Lessons Learned and Next Steps in Energy Efficiency Measurement and Attribution: Energy Savings, Net to Gross, Non-Energy Benefits, and Persistence of Energy Efficiency Behavior. http://uc-ciee.org/downloads/EEM_A.pdf.

Skumatz, L.A. (2012). What Makes a Good EUL? Analysis of Existing Estimates and Implications for New Protocols for Estimated Useful Lifetimes (EULs). International Energy Program Evaluation Conference. Rome, Italy.

Tierney J. (2011), "When Energy Efficiency Sullies the Environment", The New York Times, March 2011.

Violette, D.; Cooney, K. (December 8, 2003). Findings and Report: Retrospective Assessment Of The Northwest Energy Efficiency Alliance. Prepared for Northwest Energy Efficiency Alliance Ad Hoc Retrospective Committee. Summit Blue/Navigant Consulting. www.theboc.info/pdf/Eval-BOC_SummittBlue_NEEA_2003.pdf.

5 Resources

Ahmad, M.; Deng, A.; Spoor, S.; Usabiaga, M.; Zhao, I. (2011). *Persistence of Energy Savings in Industrial Retrocommissioning Projects*. ACEEE Summer Study on Energy Efficiency in Industry. www.aceee.org/files/proceedings/2011/data/papers/0085-000028.pdf.

Decision Sciences Research Associates. (1999). *1994 Commercial CFL Manufacturers' Rebate Persistence Study.* Report ID 529D. Pasadena: Decision Sciences Research Associates.

KEMA. (April 24, 2006). *2005 Smart Thermostat Program Impact Evaluation*. Final report prepared for San Diego Gas and Electric Company. http://sites.energetics.com/madri/toolbox/pdfs/pricing/kema_2006_sdge_smart_thermostat.pdf.

Navigant Consulting, Inc. (2010). *Kaizen Blitz Pilot. Report 1.* Prepared for Energy Trust of Oregon.

Ontario Power Authority. (2011). *EM&V Protocols and Requirement*. Ontario: 2011-2014.

Pacific Gas & Electric. (1999). *Commercial Lighting Study.* CALMAC. pp. 4-14.

Proctor Engineering. (1998). *Statewide Measure Performance study #2: An Assessment of Relative Technical Degradation Rates. Final Report.* For California Measurement Advisory Committee.

RLW Analytics. (1998). *SCE Non-Residential New Construction Persistence Study. Final Report.* For Southern California Edison. Study # SCE0064.01; 554; 530.

San Diego Gas & Electric. (March 2006). *1996 & 1997 Commercial Energy Efficiency Incentives Ninth Year Retention Evaluation.* www.calmac.org/publications/2006%5FPY96%26PY97%5FCEEI%5F9th%5FYear%5FRetention%5FEvaluation%2Epdf.

U.S. Environmental Protection Agency. (November 2007). *Model Energy Efficiency Program Impact Evaluation Guide.* www.epa.gov/cleanenergy/documents/suca/evaluation_guide.pdf.

6 Appendix A: Program-Specific Persistence Study Challenges and Issues[29]

Persistence studies provide useful information for making sensible energy efficiency (EE) investment decisions when the benefit/cost test of a measure is sensitive to changes in savings over time. As such, various persistence study challenges and issues should be examined regarding how energy savings are estimated (e.g., through measure and/or behavioral change). Table 3 summarizes persistence study challenges and issues by energy activity.

Table 3. Persistence Study Challenges and Issues

Program Measure or Activity	Characteristics	Persistence Study Challenges and Issues
New Installation, Retrofit, and Replace on Burnout	• Intervention occurs at the time measures are being replaced. • Savings result from the difference in energy use between the old equipment and the EE equipment. • An example is a lighting rebate program that provides incentives to participants for switching to higher-efficiency lighting measures.	• Cost of on-site data collection is high. • Impractical to wait for half of the units to fail so as to determine median survival time. • Some owners prematurely interrupt measure life for various reasons (such as dissatisfaction with new equipment) and switch back to less-efficient equipment. • Measure life estimates are based on failures. However, as there are few equipment failures in the early stages of equipment life, it is difficult to get an unbiased determination of expected useful life (EUL). • A lack of plug load sector data. • Business turnover has a strong effect on commercial measure lifetime. • When replacing equipment before the end of equipment life, the question of whether EE should be calculated by the delta of efficient equipment compared either to (1) replaced equipment, or (2) the equipment required by codes and standards. There is difficulty in predicting future standards.
Early Retirement	• Accelerates the retirement of inefficient equipment. • Savings result from load reduction due to absence of inefficient equipment. • An example is a refrigerator recycling program that gives participants an incentive for terminating the use of inefficient refrigerators.	• RUL is not well-studied, thus, it introduces uncertainties to future savings after the early retirement of the old equipment.

[29] Ms. Angie Lee and Mr. Mohit Singh-Chhabra of Navigant, Inc., developed this appendix.

Behavioral Programs		
Energy Activity	**Characteristics**	**Current Persistence Study Challenges and Issues**
Feedback[30]	• Programs that influence behavioral changes to obtain energy savings. • Savings result from behavioral changes. • An example is an informational program that tells households of their energy consumption as compared to their neighbors.	• Current standard behavior is going to change, and future standard behavior is difficult to predict. • A lack of studies on behavioral programs. • It is difficult to find an unbiased, uncontaminated control group.
Educational/Training	• Educational programs that provide customers with EE education. • Savings result from behavioral changes. • An example is a school education program.	• Current standard behavior is going to change, and future standard behavior is difficult to predict. • A lack of studies on behavioral programs.
Operation & Maintenance (O&M)	• Provides O&M best practices with low-cost/no-cost measures, such as adjusting control settings. • Savings result from improved O&M. • An example is retro-commissioning activity.	• Retro-commissioning programs typically have a short useful life[31], since most of the activities involve adjusting controls. • Operators who are unaware of the reason behind adjustments could revert back to the original settings.

[30] Navigant Consulting (2011).
[31] Ahmad et al. (2011).

Table 4. Measure and Behavioral Programs

Measure and Behavioral Programs		
Energy Activity	**Characteristics**	**Current Persistence Study Challenges and Issues**
Whole Building New Construction and Retrofit[32]	• Combination of both EE measures and O&M best practices. • Savings result from the difference in energy use between the old equipment and the EE equipment, as well as from O&M best practices over baseline behavior.	• It is difficult to separate out the effects of specific measures in a whole-building system, as most energy evaluations utilize billing analysis or building simulations to estimate whole-building savings.
Smart Thermostat[33]	• Thermostats are used to influence AC use. • Users obtain incentives for allowing the utility to adjust their thermostat set points while reserving the right to override the utility re-set. • Savings result from reduction in energy usage occurring from changes in AC use.	• A lack of persistence studies on smart thermostat programs.

[32] RLW Analytics (1998).
[33] KEMA (2006).

The following table presents candidate methods by study type—measure life, retention and degradation.

Table 5. Methodology Summary

Method	Method Description and Application	Data Requirements	Applicable Studies		
			Meas ure Life	Reten tion	Degrada tion
On-Site Equipment Installation Verification	• Verifications through an on-site inspection: (1) that equipment is in-place and operable, and (2) whether the application of the equipment has changed. • Applicable to evaluating measure programs. • An example is a measure life/EUL study of a commercial lighting incentive program using on-site audits[34].	• Measure make and model. • Spot measurements to supplement visual inspection. • Date installed and date when measure became inoperable or was removed.	x	x	
On-site Equipment Measurement and Testing	• Measurement (short term or long term) of equipment performance, focused on collecting data and ensuring equipment is use as designed. If it is not, then identifying the reasons the usage differs from the equipment's design intent.) • Applicable to evaluating measure programs. • An example is a degradation study of high-efficiency motors.	• Measure make and model. • Use of equipment as designed. • Observation of failure rates.			x
Laboratory Testing	• Measurement of energy use of both EE and standard equipment over time in unoccupied facilities. • Laboratory testing must account for the operational conditions expected for installations. • Applicable to evaluating measure programs. • An example is a degradation study comparing existing and high-efficient air compressors.	Energy use of equipment over time.			x

[34] San Diego Gas & Electric (1999).

Benchmarking and Secondary Literature Review	• Engineering review of equipment degradation and uncertainties. The literature search should include journal articles, conference proceedings, manufacturer publications, and publications of engineering societies. • Applicable to evaluating both measure and behavioral programs. • An example is an assessment of measure technical degradation rates by conducting a meta-review on secondary literature.[35]	Equipment and/or behavior degradation and uncertainties.	x	x	x
Telephone Surveys/ Interviews	• Interviews of program participants about: (1) their consumption patterns compared to EE equipments' design intent, and (2) whether the EE equipment is in place and operable. • Applicable to evaluating both measure and behavioral programs. • An example is a persistent study of an O&M program studying behavioral retention.[36]	Equipment failures and/or replacement behavior, including time of failure and/or replacement, and the number of failures and/or replacements.	x	x	x
Billing Analyses – Fixed Effects and Statistically Adjusted Engineering Models[37]	• Statistical analysis to model the difference between customers' energy usage pre- and post-analysis periods, using real customer billing data over multiple years. • Applicable to measure and behavioral programs. • An example is evaluating multiyear savings persistence on commercial lighting technologies.[38]	Customer billing data over time.		x	x

[35] Proctor Engineering (1998).
[36] Navigant Consulting, Inc. (2010).
[37] Pacific Gas & Electric (1999).
[38] Quantum Consulting (1998).

Survival Curves	• Linear, logistics, exponential, or hazard models estimating equipment survival rate. The model choice depends on equipment characteristics and previous research. • Applicable to measure and behavioral programs. • An example is estimating the EUL of equipment installed in a new construction project using survivor function and hazard function.	Independence of equipment failure and EUL.	x		
Controlled Experiment	• Experiment developed across census, randomly assigning participants into treatment and control groups. • Applicable to behavioral programs. • An example is a retention study of a behavioral energy program over multiple years.	Customer billing data of control group and treatment group over time.		x	x

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter author wishes to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Hossein Haeri of The Cadmus Group
- Brent Barkett of Navigant.

In addition, the chapter author wishes to acknowledge helpful comments submitted through the Stakeholder Review process.

Suggested Citation

Violette, D.M. (2017). *Chapter 13: Assessing Persistence and Other Evaluation Issues Cross-Cutting Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68569. http://www.nrel.gov/docs/fy17osti/68569.pdf

Acronyms

AC	air-conditioning
Btu	British thermal unit
CALMAC	California Measurement Advisory Council
CFL	compact fluorescent lamp
DOE	U.S. Department of Energy
EM&V	evaluation, measurement, and verification
EMS	energy management system
EUL	effective useful life
FCM	forward capacity market
HVAC	heating, ventilating, and air conditioning
ISO-NE	Independent System Operator-New England
M&V	measurement and verification
NPV	net present value
RUL	remaining useful life
SAS	Statistical Analytics Software
TDF	technical degradation factor
UMP	Uniform Methods Project

Table of Contents

List of Figures

List of Tables

1 Introduction

Addressing other evaluation issues that have been raised in the context of energy efficiency programs, this chapter focuses on methods used to address the persistence of energy savings, which is an important input to the benefit/cost analysis of energy efficiency programs and portfolios. In addition to discussing "persistence" (which refers to the stream of benefits over time from an energy efficiency measure or program), this chapter provides a summary treatment of these issues:

- Synergies across programs
- Rebound
- Dual baselines
- Errors in variables (the measurement and/or accuracy of input variables to the evaluation).

This first section of this chapter contains a definition of persistence and identifies issues in its evaluation. The state of the practice in persistence is addressed, examples taken from persistence studies are presented, and recommendations for addressing persistence are presented at the end of the section. The other evaluation issues are addressed in the second section of the chapter. Appendix A presents a matrix of persistence issues and methods by program type.[1]

[1] As discussed in the section "Considering Resource Constraints" of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

2 Persistence of Energy Savings

Understanding persistence is critical to making good decisions regarding energy efficiency investments, so this section outlines program evaluation methods that can be employed to assess persistence—the reliability of savings over time. Energy efficiency program benefits are measured as the net present value (NPV) of a stream of benefits based on the energy and demand savings[2] achieved by the program. Depending on the mix of measures and their assumed lives, these benefits may extend to 15 years (or more) for some measures. As a result, assumptions about the persistence of savings over time influence the energy efficiency benefit-cost tests. Extrapolating savings beyond the evaluation period has often been based on engineering judgment, manufacturer specifications, and some empirical work (the factors used to develop projections of measure lifetimes and degradation).

The protocols developed under the Uniform Methods Project (UMP) in other chapters generally focus on estimating first-year savings. There is also some discussion, however, about estimating first- and second-year savings when more participants from a second program year are needed for the impact evaluation. These initial evaluations are often quite detailed, assessing both the savings and the quality of the program in terms of installation, engineering calculations, and equipment selection (where on-site visits are used to validate initial "claimed" estimates).

2.1 Addressing Persistence

Persistence of savings encompasses both the retention and the performance degradation of measures. Together, these factors are used to estimate how the *claimed* persistence values used in program planning can be updated based on *evaluated* savings values.[3] Different jurisdictions define and treat the components of overall persistence differently. As a result, defining what is meant by overall persistence and addressing some of the subtle context issues are important to the discussion.

There are a number of subtle aspects to the context and definition of overall persistence. Consistent and practical definitions for use in developing estimates of the overall persistence of savings over time were developed for the Joint Massachusetts Utilities (Energy and Resource Solutions 2005).[4] In that study, overall persistence is divided into two components: (1) measure life and (2) savings persistence.

Recognizing that definitions for *persistence* and *realization of savings* are not nationally consistent, the definitions based on the Massachusetts framework and outlined below provide a structure that can be addressed by evaluation and verification methods. That is, these definitions

---

[2] This chapter focuses on estimating energy savings, but the persistence of reductions in demand may also be important for some measures and programs. Issues raised here may also be important for programs and policies focused on reducing demand during peak periods.

[3] In this chapter and consistent with other chapters, *claimed* savings means the same as *ex ante* savings and *evaluated* savings is used instead of *ex post* savings. This note is to eliminate confusion for those more familiar with the use of "*ex ante"* (initial savings estimates) and "*ex post"* (evaluated savings) terminology in describing evaluation methods.

[4] This study for the Joint Massachusetts Utilities' defines "measure life" as the median number of years that a measure is installed and operational. This definition implicitly includes equipment life and *measure* persistence. However, *savings* persistence is the percentage of change in expected savings due to changed operating hours, changed process operation, and/or degradation in equipment efficiency relative to the baseline efficiency option.

use categories of effects and factors that can be quantified using evaluation methods. For example, it is difficult to estimate technical measure life based on on-site inspections, as there may be many reasons that a measure is no longer in place. Thus, technical measure life and other reasons for measure non-retention are combined in the definition "measure life," which is simply the time a measure can be expected to be in place and operable.

2.1.1 Definitions

The definitions of key terms used in this chapter are these.

2.1.1.1 Measure Life or Effective Useful Life

This is the median number of years that a measure is in place and operational after installation. This definition implicitly includes equipment life and measure persistence (defined below), but not savings persistence.

- "Equipment life" is the number of years installed equipment will operate before it fails.
- "Measure persistence" takes into account business turnover, early retirement or failure of the installed equipment, and any other reason the measure would be removed or discontinued.

2.1.1.2 Savings Persistence

This is the percentage of change in expected savings due to changed operating hours, changed process operations, and/or the performance degradation of equipment efficiency relative to the baseline efficiency option. For example, an industrial plant that reduces operation from two shifts to one shift may then have a savings persistence factor of 50%, as only half of the projected energy savings would be realized. Also, improper operation of the equipment may negatively affect *savings* persistence, so training and commissioning could improve savings persistence. Finally, most equipment efficiency degrades over time, so annual energy savings may increase or decrease relative to the efficiency degradation of the baseline efficiency option.

Figure 1 illustrates how the two persistence factors are used to produce savings that are adjusted for persistence: Savings Adjusted for Persistence = (Measure Life Factor) x (Savings Persistence Factor) x (Initial Savings Estimate).

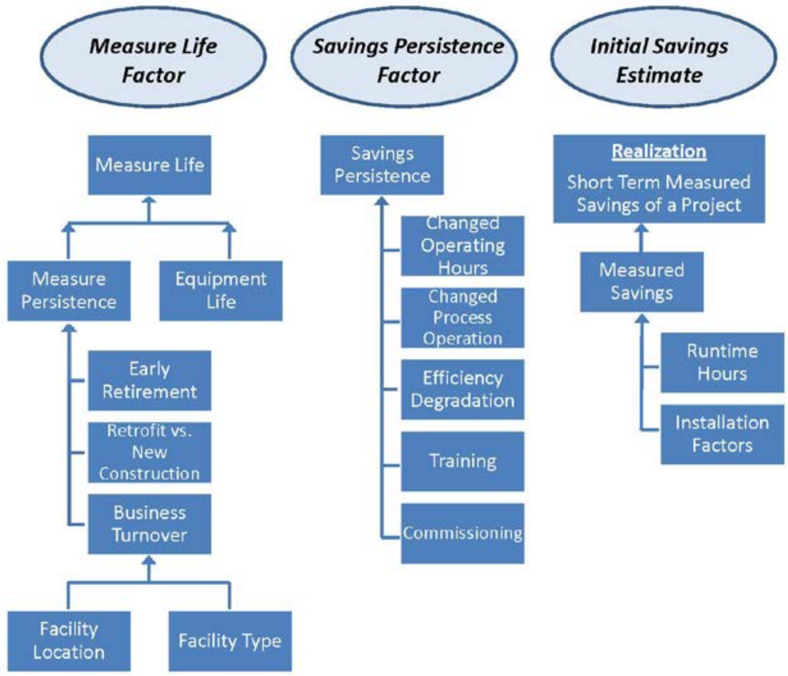


Figure 1. Relationship of measure life, savings persistence, and initial savings estimates[5]

2.1.2 Factors for Selecting a Persistence Study

The following are several important factors to consider when selecting the type of study to examine energy savings persistence.

2.1.2.1 Available Claimed Estimates of Persistence

There are almost always initial *claimed* estimates of the assumed stream of savings for a program (based on current estimates of measure life and degradation). These estimates are used in the initial benefit/cost analyses conducted as part of program design or in the benefit/cost tests of initial program evaluations efforts. As a result, most studies of persistence test the initial claimed stream of savings against the evaluated results to check for significant differences.[6] The outcome is often presented as a realization rate (that is, the *evaluated* values divided by the initial claimed values), which is the year-by-year savings estimate used in benefit/cost studies.

2.1.2.2 Uncertainty in Claimed Estimates

When deciding whether to conduct a new study of persistence—and the corresponding level of effort required—consider the confidence that the evaluator or decision-maker has in the claimed stream of savings values. If the uncertainty is perceived as being high *and* a sensitivity analysis shows that plausible revisions to persistence of energy savings substantively changes the results

[5] Source: Adapted from Energy and Resource Solutions (2005).
[6] Starting with a set of claimed savings allows for the use of evaluation methods that leverage these initial data through the use of ratio estimates and a "realization rate" framework.

of benefit/cost tests, then a new study may be worthwhile. Such an undertaking regarding persistence may result in revisions to the current *claimed* estimates.

For example, measures that account for greater savings, have shorter measure life values, or may be subject to near-term degradation in savings are more important to evaluate, as they will have a greater impact on the resulting benefit/cost tests. However, changes in measure life that do not take effect until the 14th or 15th year of the measure may be discounted in the NPV calculation (discussed below). Thus, in terms of the effect on the benefit/cost calculation, the additional work needed to estimate these values may not be worthwhile.

2.1.2.3 Discounting Values of Energy Savings Over the Life of the Measure

The stream of program benefits over time is discounted, resulting in near-term savings estimates that have a larger impact on the NPV of benefits than the values further out in the future. For example, the effect of research on the measure life of a second refrigerator retirement that extends it from six years to eight years would be muted somewhat in the benefit/cost analysis due to discounting. Specifically, the energy savings from this updated measure life of two additional years would be muted in its application by discounting the benefits for year seven and year eight. The impact of discounting depends on the discount rate being used and the measure life.[7]

2.1.2.4 Differences in Baseline and Energy Efficiency Energy Streams of Benefits

Energy savings calculations are based on the difference between the post energy efficiency state and the assumed baseline. If the baseline equipment has the same level of degradation in performance, then the energy savings factor due to degradation would be 1.0 *and* it would be appropriate to assume constant energy savings over the life of the energy efficiency measure.[8] In fact, if the relative persistence of savings is higher for the energy efficiency measures compared to a baseline consisting of standard measures, then energy savings not only persists, but can increase over time.

[7] For example, if a discount rate of 5% is used, the savings will be reduced by 0.78 multiplied by the energy savings at five years. At 10 years and a 5% discount rate, the new value would be 0.61 multiplied by the energy savings. At a discount rate of 7% for a 10-year period, the value would be 0.51 multiplied by the energy savings.

[8] The report from Peterson et al. (Peterson et al. 1999) is a good example of degradation being measured for both an efficient appliance offered by an energy efficiency program and standard equipment. This study showed that the high-efficiency coils start with and maintain a higher efficiency than standard efficiency coils. The slower degradation rate increases the life of the equipment, and the equipment uses less energy over its operational lifetime. Even though both high-efficiency units and standard units showed performance degradation over time, the lower rate of degradation in the high-efficiency units resulted in a recommended degradation factor exceeding 1.0 in most years. This factor increased from 1.0 to 1.08 over the 20-year expected life of the unit, indicating that savings not only persisted, but actually increased relative to the baseline over the assumed life of the equipment.

These four factors are meant to address the following questions:

- If a persistence study is conducted, is there a reasonable likelihood that the new trend in energy savings over time would be substantively different from the assumptions used in the initial benefit/cost analyses?
- Would the NPV benefits of the program change with a new persistence factor, the discount rate being used, and the likely change in the baseline energy use level that may also be due to performance issues of the baseline equipment?

There may be good reasons to assess persistence, as many factors can influence the stream of energy savings over a three- to 10-year period. The most common of these factors are listed in Table 1.

Table 1. Factors Influencing Persistence

Residential Sector Programs and Measures	Commercial and Industrial Sector Programs and Measures
1. Changes in ownership	1. Business turnover
2. Maintenance practices	2. Remodeling
3. Changes in equipment use	3. Varying maintenance
4. Behavioral changes	4. Operating hours and conditions
5. Occupancy changes	5. Inappropriate installation of equipment
6. Inappropriate installation of equipment	6. Manufacturer performance estimates that do not reflect in-field operating conditions.
7. Manufacturer performance estimates that do not reflect in-field operating conditions.	

Sensitivity analyses using the benefit/cost models can highlight those measures for which adjustments in persistence will have the largest impact. This information can then be used to prioritize persistence evaluation efforts. Thus, before deciding whether additional analyses are needed, test the sensitivity of NPV benefits to potential changes in the persistence of savings. This can help determine whether the impact may be large enough to merit a substantial study effort, or sufficiently small, requiring only a modest retention study.

2.2 State of the Practice in Assessing Persistence

Professional judgment plays a significant role in selecting a method for assessing persistence. The *California Energy Efficiency Evaluation Protocols* (CPUC 2006) has several types of retention, degradation, and measure life/effective useful life (EUL) studies from which to select, based on the priority given to the issue by regulatory staff or other stakeholders.

Evaluators seem to rely on the following two processes for developing estimates of persistence:

- ***Database or Benchmarking Approach.*** This entails developing and regularly updating[9] a database of information on measure life and performance degradation.
- ***Periodic In-Field Studies.*** This entails performing selected in-field studies of program participants from earlier years.

These two approaches are not necessarily mutually exclusive. The database/benchmarking approach is often used when (1) there are a large number of energy efficiency measures, (2) there are concerns about the sample sizes required for in-field studies, and (3) the cost of conducting in-field persistence studies is an issue. Periodic studies may be used for updating a database of measure life and performance degradation. Such studies are also useful when focusing only on those measures that account for a large fraction of the savings. Additionally, in-field studies of program participants that are conducted a number of years after participation provide direct information on persistence of savings for that program.

2.3 Database/Benchmarking Approaches

The three examples of database/benchmarking approaches presented below are based on:

- Engineering judgment
- Experience with the energy efficiency measures
- Information on local and regional conditions to develop tables of measure lives for use in energy efficiency program planning.

These values are often used as deemed values for persistence and applied to produce estimates of the energy savings over time (as inputs to benefit/cost calculations). An assessment of this approach follows the examples. (References to each study are provided for those wanting more information on the methods used beyond the short descriptions provided below.)

2.3.1 Example Study 1: GDS Associates (GDS Associates 2007)

Objective: The measure life values presented in this report were developed to meet the following conditions:

- Accurately reflect conditions for measures installed by energy efficiency programs in the New England states that have supported this research effort
- Satisfy any Independent System Operator-New England (ISO-NE) requirements (for example, for definition and documentation sources)
- Work as common values, accepted by all New England states for the forward capacity market (FCM) (that is, the ISO-NE forward capacity market).

[9] As it is important that these benchmarking studies be updated on a regular basis, the cost of these updates should be included in the cost estimate for using this approach. While these studies may not appear costly on a one-time basis, the effort required to update the database regularly can be significant. This is important, as these databases are sometimes the source of deemed values for measure life and persistence of savings used in evaluation efforts.

Methodology: "Reviewed all secondary data collected and developed a preliminary list of potentially applicable residential and C&I [commercial and industrial] measures. This list was then distributed to program administrator staff within the SPWG [State Program Working Group] for review and to obtain additional program-specific measure life values and associated documentation sources. GDS compiled all responses and developed initial measure life recommendations for SPWG member consideration."

2.3.2 Example Study 2: KEMA (KEMA 2009)

Objective and Methodology: "The principal objective of this study was to update the current measure life estimates used by the Focus Evaluation Team and the Focus Program. **The evaluation team's approach to this study consisted entirely of secondary research**; the team did not conduct primary research, fieldwork, or produce a savings persistence study." (Emphasis added.)

2.3.3 Example Study 3: Energy and Resource Solutions (ERS 2005)

Objective: "The primary goals of the Common Measure Life Study were as follows:

- Define measure life and related terms, such as persistence
- Review the provided table of current measure lives
- Survey other utility energy efficiency programs
- Develop a table of technological measure lives
- Recommend common measure lives and persistence assumptions to be used by the sponsors."

Methodology: "ERS [Energy and Resource Solutions] reviewed the tables of agreed-upon and disputed measure lives provided by the sponsoring utilities. As tasked in our proposal, we researched several sources to use in support of selecting individual measure lives. We first thoroughly researched the CALMAC [California Measurement Advisory Council] database. The CALMAC database provides a public depository for all persistence, technical degradation factor (TDF) and other related studies performed in the State of California. Next, we surveyed many electric utilities and state utility commissions throughout the nation, obtaining other utilities' tables of measure lives. We obtained measure life tables used in 8 states by at least 14 different utilities. Finally, we performed a literature search, referenced technical sources and consulted equipment manufacturers to establish a table of technical lives for each measure. In conjunction with these efforts, we specifically researched the effect of New Construction versus Retrofit status on measure lives, as well as the effect of Small versus Large businesses."

2.4 The Challenges of New Technologies and Measures

The methods in the three examples above have produced useful estimates for a wide number of measures where practical information exists from measure installations and fieldwork. However, new technologies and measures installed less frequently pose greater challenges for this judgment-based benchmarking approach. For many widely implemented energy efficiency measures, both the evaluation work and additional on-site engineering work (such as installation and maintenance) provide a basis for the use of informed engineering judgment. A series of retention/survival rate studies in California—conducted from 1994 to 2006—found that most *claimed* estimates could not be rejected by the in-field studies. However, the in-field studies often had small sample sizes for certain measures and short time frames that did not allow for many failures to occur in the dataset.

Some important measures in these engineering and expert-developed measure life tables may not have fared well. Both residential lighting and commercial lighting have provided a large fraction of savings, and the persistence of these savings has been controversial. Nexus (2008) found that the life for certain lighting measures depends not only on the equipment, but also on the program design.

Skumatz (Skumatz et al. 2009) (Skumatz 2012) critiques the database/benchmarking approach, which is based on engineering judgment combined with literature reviews. Skumatz (2012) identifies strengths and weaknesses in this approach compared to on-site data collection, and she offers suggestions for improving current estimates. Skumatz notes that measure life values existing in tables often vary by more than 25%, and that this has "precisely the same impact on a measure's or program cost-benefit ratio" as savings values that are off by 25%.

While this comment has merit, the measure life and persistence factors will start at 1.0 in the initial years of the program and then gradually change. This change in savings is offset to some degree by the discounting of benefits from five, 10, and 15 years out. Also, this single measure with varying measure life values across engineering-based tables may not represent the composite effective life of a group of measures that make up a program.

2.5 In-Field Persistence Studies (Survey and On-Site Data Approaches)

Methods that make use of in-field data collected on program participants at some point after they participated in an energy efficiency program generally rely on:
Surveys or on-site visits to determine whether the measure is still in place and operable, or, if the measure was removed, when and why[10]

- Statistical analyses using regression-based methods to estimate retention/survival models that produce estimates of the survival or failure rates of energy efficiency measures.

The *California Energy Efficiency Evaluation Protocols*[11] specified these three categories of methods used for in-field studies of persistence:

- ***Retention Studies*** provide the percentage of the measures that are in place and operable at a point in time. Retention studies identify technology design, define operable conditions, and describe how operable conditions could be measured.

[10] One reviewer suggested that the surveys referred to in this section should specifically include online approaches. The topics of using online surveys to obtain customer-specific information and combining online surveys with other methods are discussed in the "Survey Research" chapter.
[11] The methodology language from the *California Energy Efficiency Evaluation Protocols* (California Public Utilities Commission 2006) has been adapted to fit the measure life definition and persistence structure used in this chapter. One difference is the use of *persistence* as the overarching term for all types of changes in energy savings over time, which the California Protocols document addresses in the "Effective Useful Life Protocol" section (p. 105). The California *Protocols* still contain the most comprehensive discussion of methods for assessing persistence.

- ***Measure Life/EUL*** estimates the median numbers of years that the measures installed under the program are still in place and operable. This value is calculated by estimating the amount of time until half of the units will no longer be in place and operable.
- ***Performance Degradation*** uses both technical and behavioral components to measure time-related and use-related changes in energy savings relative to a standard efficiency measure or practice. In general, both standard equipment and energy efficiency equipment become less efficient over time, regardless of the equipment measure life. This factor is a ratio reflecting the decrease in savings due to performance degradation from the initial year savings.

2.5.1 Retention and Measure Life Studies

A retention study determines the number of installed and operable measures at a given point in time. A measure life study is an extension of a retention study, where there is adequate data to allow for the development of a statistical model (commonly called a "survival analysis") to estimate failures that might occur after the data are measured.

Information from the retention model provides an estimate of the measures that were installed and operating at a point in time, which allows the evaluator to calibrate the *claimed* savings and produce adjusted *evaluated* estimates of savings over time. The current estimates of persistence are adjusted to account for the new information *and* the stream of savings over the year. These estimates could, for example, be adjusted in year four to be consistent with the retention study. This ratio for year four would then be used to adjust the savings in all subsequent years.

The measure life estimation methods, which are based on survival analysis, provide more information. However, estimating measure life requires a much larger sample—one that contains an adequate number of both installed and missing (that is, uninstalled or replaced) equipment.

The following are two types of retention and measure life methods, which have been used to estimate the survival models that produce estimates of measure life. (Studies using these methods are described later in this section.)

2.5.1.1 In-Place and Operable Status Assessment (Using On-Site Inspections)

The in-place assessment studies are verified through on-site inspections of facilities. Typically, the measure, make, and model number data are collected and compared to participant program records, as applicable. As-built construction documents may also be used to verify selected measures when access is difficult or impossible (such as wall insulation). Spot measurements may be used to supplement visual inspections—such as solar transmission measurements and low e-coating detection instruments—to verify the optical properties of windows and glazing systems.

Correct measure operation is observed and compared to the project's design intent. Often, this observation is a simple test of whether the equipment is running or can be turned on. However, the observation and comparison can extend to changes in application or sector, such that the operational nature of the equipment no longer meets the design intent. For example, working gas-cooking equipment that had been installed in a restaurant but is now installed in the restaurant owner's home is most likely no longer generating the expected energy savings, so it would not be counted as a program-induced operable condition.[12]

2.5.1.2 Non-Site Methods

Typical non-site methods include telephone surveys/interviews, analysis of consumption data, or the use of other data (such as from energy management systems). The goal is to obtain essentially the same data as would be gotten through an on-site verification; however, there is the potential for collecting inaccurate data, due to a number of factors (discussed in Chapter 11: *Sample Design*).

2.5.1.3 Examples of Retention and Measure Life Studies

Two examples of these types of studies were performed by KEMA and by Nexus Market Research.

- KEMA (KEMA 2004) used a telephone survey to gather information on refrigerators at years four and nine as part of a review of an appliance recycling program.
- Nexus Market Research (Nexus Market Research 2008) used on-site verification data to conduct a measure life study of residential lighting measures.

Both studies provide good examples of collecting information for a basic retention study, and they serve as illustrations of the statistics necessary to estimate a survival model (Allison 1995).[13] Each is discussed below.

Example Study 1: KEMA (KEMA 2004). Conducted with program participants from the years 1994 through 1997, this study looked at retained savings over this period.

For each year, the measure life/EUL estimate reflects the following factors:

- The time at which half of the recycled appliances are from participating premises that have added an appliance
- The time at which half of the recycled appliances would have been out of service without the program influence.

The KEMA study illustrates one way in which the *claimed* and *evaluated* measure life values can be used. As stated in the study:

---

[12] In addition to this language, the *California Energy Efficiency Evaluation Protocols* outlines certain sampling criteria that must be met in California. However, these criteria may vary in accordance with the requirements of different jurisdictions.

[13] To assist evaluators, the *California Energy Efficiency Evaluation Protocols* states: "Multiple statistical modeling packages (SAS, Stata, SPSS, R, S+, and others) provide survival analysis programs. There are several commercial and graduate textbooks in biostatistics that are excellent references for classic survival analysis. One of these used as reference for some of the prior EUL studies in California is the SAS statistical package and the reference *Survival Analysis Using the SAS System: A Practical Guide* by Dr. Paul D. Allison, SAS Institute, 1995. Several model functional forms are available and should be considered for testing. These forms include logistic, logistic with duration squared (to fit expected pattern of inflection point slowing of retention losses), log normal, exponential, Weibull, and gamma."

> For each of the program years from 1994 through 1997, both refrigerators and freezers have a claimed (or *ex ante*) estimate of measure life/EUL of six years, which has been used in the earnings claims to date. A measure's evaluated measure life/EUL is the value estimated by a persistence study. If a measure's claimed measure life/EUL is outside the 80% confidence interval, the measure's evaluated measure life/EUL may be used for future earnings claims. Otherwise, the measures claimed value will continue to be used in earnings claims.

Figure 2 is a replication of Table E-1 from the KEMA study, which shows the comparison between the *claimed* and *evaluated* measure life/EUL estimates. In this case, the measure life results showed that the program was underestimating the measure life/EUL values *and* that the realization rate exceeds 1.0.

Table E-1
1994-1997 Appliance Recycling Program
Summary of Effective Useful Life Estimates

			EUL (years)					
						80% Confidence Interval		
Program Year	Measure	End Use	Ex Ante	Ex Post (estimated from study)	Adopted ex post (to be used in claim) Lower Bound	Lower Bound	Upper Bound	EUL Realization Rate (adopted ex post/ex ante)
1994	Freezer	Refrigeration	6.0	8.0	8.0	8.0	11.0	1.33
	Refrigerator		6.0	8.0	8.0	8.0	11.0	1.33
1995	Freezer	Refrigeration	6.0	8.0	8.0	8.0	11.0	1.33
	Refrigerator		6.0	8.0	8.0	8.0	11.0	1.33
1996	Freezer	Refrigeration	6.0	8.0	8.0	8.0	8.0	1.33
	Refrigerator		6.0	8.0	8.0	8.0	8.0	1.33
1997	Freezer	Refrigeration	6.0	8.0	8.0	8.0	8.0	1.33
	Refrigerator		6.0	8.0	8.0	8.0	8.0	1.33

Figure 2. KEMA (2004) Table E-1

Example Study 2: Nexus Market Research (2008). This study examined the measure life of lighting products distributed through energy efficiency programs in New England.

The definition of measure life is the same as presented above in *Addressing Persistence* and used in the Energy and Research Solutions (2005) example application presented above. Specifically, Nexus states that:

> [T]he measure life estimates do not distinguish between equipment life and measure persistence; our estimates—one for each measure category—include both those products that were installed and operated until failure (that is, equipment life) as well as those that were retired early and permanently removed from service for any reason, be it early failure, breakage, or the respondent not liking the product (that is, measure persistence).

Nexus drew a random sample of participants based on the type and number of products they had obtained through the programs. The report states, "We collectively refer to these sample products as the 'measure life products.'"

Auditors visited 285 homes to inventory lighting products, and Nexus designed a respondent survey to learn more about the measure life products and other lighting products found in the home. These survival analyses were based on the following methods and, ultimately, Nexus used estimates resulting from Method 3.

- Method 1: Measure Life Tables
- Method 2: Logit Regression
- Method 3: Parametric Regression Models of Survival Analysis.

The results showed that the measure life for compact fluorescents (CFLs) varies by program design (that is, whether the program was coupon-based, direct install, or a markdown at a retail facility). The results of the Nexus (2008) study are shown in Table 2.

Table 2: Nexus (2008) "Recommended Estimates of Measure Life—Decimals"

Product	Measure Life	80% Confidence Interval	
		Low	High
Coupon CFLs	5.48	5.06	5.91
Direct Install CFLs	6.67	5.97	7.36
Markdown CFLs (all states)	6.82	6.15	7.44
Coupon and Direct Install Exterior Fixtures	5.47	5.00	5.93
Markdown Exterior Fixtures	5.88	5.24	6.52
All Interior Fixtures	Continue using current estimates of measure life		

Nexus deemed a representation of the results—at an 80% confidence interval—as being accurate enough for the purposes of this study. Nexus recommended measure life estimates for three measures: one for compact fluorescent lamps (CFLs; coupon, direct install, and markdown)[14] and two for exterior fixtures (markdown and all other programs).

Nexus did not recommend an estimate of measure life for interior fixtures, as the timing was too early in the measure lifecycle to provide a reliable estimate. This occurs with a number of measure life studies that are conducted too early (before there have been enough failures or un-installs to allow for statistical modeling of measure life).

2.5.2 Examples of Degradation Studies

While there are few reports that directly focus on the degradation of savings, two types of studies are available, and they are described below:

- Focusing on technical degradation (one of the clearest examples is by Proctor Engineering in 1999 [Proctor Engineering 1999])

[14] Due to the diversity of program types throughout the region, Nexus used the term "markdown" to refer to both markdown programs (offered in all of the states) and buy-down programs (offered in some of the states).

- Performing billing analyses at some point after participation to capture all of the factors that impacted persistence of savings. (In 2011, Navigant performed a billing analysis of a customer information program, which was used to examine persistence of impacts across two years for a behavioral program. [Navigant 2011])

Example Study 1: Proctor Engineering (Proctor Engineering 1999). The purpose of this project was "to examine the relative technical degradation of demand side management (DSM) measures compared to standard efficiency equipment. This project covers two major DSM measures: commercial direct expansion air conditioners (Comm. [direct expansion] DX AC) and EMS [energy management systems]."

Proctor Engineering's methodology involved establishing a time-series estimate—derived from available research—for condenser and evaporator coil fouling rates. Proctor used laboratory testing to modify the estimated fouling rates and establish a profile for coil fouling. It tested both high-efficiency and standard efficiency coils in a controlled laboratory environment, and both were subjected to continuous fouling. Proctor then monitored the efficiency of the air conditioner at various intervals to document the effects.

This study found that (1) the impact on standard equipment was greater and (2) the high-efficiency units actually had a higher level of savings persistence. The end result was that "testing shows that the TDF [technical degradation factor] for this measure is greater than one." This is an example of degradation needing to be conducted with reference to standard efficiency equipment. Energy efficiency measures may have performance degradation, but so does standard equipment. If the energy efficiency measures have a lower rate of degradation, then savings increase (as measured against the standard equipment baseline).

To assess EMS, Proctor used an on-site methodology rather than laboratory testing. The research data showed that although there is some EMS savings degradation at some locations, other locations show increasing savings. Some of the causes for this persistence are:

- No instances of disconnected or non-operational EMSs were found.
- The vast majority of EMSs appeared to be operated in a competent and professional manner.
- EMS operators had found that the EMS was a useful tool in performance of their jobs.

Proctor Engineering contrasted its work with other EMS studies showing greater degradation due to operational issues. Proctor explained the comparatively high level of persistence it found as being due to the high interest of the program participants in saving energy. The more random group of facilities in the comparison may not have been involved in EMS-related energy efficiency programs.

Proctor also conducted a billing analysis to confirm these findings. For this billing analysis, it combined the consumption data from all of the sites and then estimated the persistence of savings over time. The regression process provided statistically significant estimations at the 95% level.[15]

The primary purpose of this research was to establish the TDFs, estimated for each measure. The results from Proctor's study, seen in Figure 3, shows that the degradation factors are greater than 1.0 for the high-efficiency DX AC equipment. This indicates the degradation was less for the high-efficiency DX AC equipment than for the standard efficiency equipment.

Table ES-1 TDF

Year	EMS	Comm DX AC
1	1.00	1.00
2	1.00	1.00
3	1.00	1.00
4	1.00	1.01
5	1.00	1.01
6	1.00	1.01
7	1.00	1.01
8	1.00	1.01
9	1.00	1.01
10	1.00	1.02
11	1.00	1.02
12	1.00	1.02
13	1.00	1.02
14	1.00	1.02
15	1.00	1.02
16	1.00	1.02
17	1.00	1.02
18	1.00	1.02
19	1.00	1.06
20	1.00	1.08

Figure 3. Proctor Engineering (1999) Table ES-1

Still, the difference is small through year 18, and this size of effect might not show up in benefit/cost analyses due to the discounting required to obtain an NPV of savings benefits.

Example Study 2: Navigant (Navigant 2011). This study examined the short-term persistence of a behavioral information program using billing data across multiple years, as short-term persistence may be an important factor for these programs.

The program was designed to assist and encourage customers to use less energy. These types of programs are increasing in the industry; for example, OPOWER, Inc., offers residential customers regular Home Electricity Reports about their electricity consumption to help those customers manage their electricity. In combination with other information, these reports compare

[15] References to statistically significant results in regression analyses must be carefully interpreted. The analysis may have been a test to determine if the effect was significantly different from zero (±100% precision). Alternatively, the test may have actually established a precision level of ±10% or another level of precision, (for example, 30%). A statement of statistically significant results should be accompanied by an explanation for interpreting that statement in terms of the level of precision being used in the test of significance.

a household's electricity use to that of its neighbors and then suggest actions to reduce electricity use. It is hypothesized that presenting energy use in this comparative fashion creates a social nudge that induces households to reduce their consumption.

Navigant evaluated the first 29 months of the program, with an emphasis on the second program year. The following main research questions were addressed in the evaluation and presented in this report:

- Does the program continue to generate savings?
- What is the trend in program savings? Is there a ramp-up period to savings? If so, for how long? Are savings now relatively stable, increasing, or falling?
- Do program savings increase with usage?

The evaluation of this program entailed developing a random control group and conducting a fixed-effects regression analysis, which is a common evaluation method. This regression method is discussed in the "Whole House Retrofit" chapter of this UMP report.

Navigant's results showed that the effects of slightly more than 2% of the energy savings persisted across the 29 months examined in the study, after an initial ramp-up period of approximately 10 to 12 months. The small effect size required a large sample of customers for the regression analysis to produce reliable results. For this behavioral program evaluation, there were more than 20,000 treatment customers and a control group of more than 30,000 customers. Thus, large samples are needed to identify small effect sizes from energy efficiency programs.

This regression framework can be applied to a third and fourth year of data to assess longer-term participation.

2.6 Persistence Recommendations and Conclusions

Evaluators address the issue of persistence of savings from energy efficiency programs because of the impact that the stream of savings estimates has on the benefit/cost tests of measures and programs. While some measure life values are estimated at more than 20 years, most benefit/cost assessments are estimated out at least 10 years or, more commonly, 15 to 20 years.

The approaches discussed in this chapter include methods to address measure life and savings performance, which may be impacted by operating conditions, behavioral changes, turnover in building occupancy, changes in measure use, and other factors. To date, the tools and methods that make up the recommended tool kit for evaluators include:

- Benchmarking and database development for measure life values and savings persistence
- On-site analyses of equipment
- Survey methods for select measures amenable to survey techniques
- Single-year estimations of equipment retention and operation
- Multiyear statistical analyses based on survival models

- Technical degradation studies based on engineering review
- Technical degradation based on laboratory testing
- Billing analyses that capture overall persistence (that is, that assess savings directly and capture all changes in savings for the time period being analyzed).

The review of methods illustrates the different ways persistence can be addressed. Research is continuing in this area, and methods have been adopted in different jurisdictions. As with any area of evaluation, there will always be improvements. The *Appendix* to this chapter presents tables outlining program and measure persistence study challenges and issues.

The balance of this section presents practical recommendations for assessing the persistence of savings. The goal of evaluation is to help stakeholders make good decisions about investments in energy efficiency programs, and this requires both an understanding of the techniques and applied judgment.

2.6.1 Recommendations

1. ***Before determining whether to undertake a large-scale persistence study of a program or measure (or even to undertake such a study at all), consider whether the results of the study are likely to have a material impact on the economics of the program.*** Persistence of savings refers to the stream of savings expected from a measure or program over a period of years. If the study's revised persistence of savings is expected to be small and to occur 10 or more years or more in the future, then the impact of that change may not have a large effect on the cost-benefit economics.

   Keep these considerations in mind when deciding:

- Benefit-cost tests are based on NPVs that discount the streams of benefits and costs. A change in measure life by a year or two *and* changes for long-lived measures may not have much impact after they are discounted.
- The performance degradation of energy efficiency measures should be assessed relative to that of the standard efficiency equipment, as both will have performance degradation. The difference between these two values determines the impact on savings.

2. ***Select the methodology that best fits the individual circumstances of the measure/program being evaluated.***

- Pick the method most appropriate to the magnitude of the effect expected. Before conducting the study, take a forward-looking view of what might be learned. While this may seem difficult, researchers across the evaluation community and the industry make these decisions on a regular basis. The key is to ensure that the information produced is worth the effort expended to produce it. The goal is to obtain information that decision makers need for making good decisions regarding energy efficiency investments.
- Measures that may have persistence impacts within the first three to seven years are the most important to study because of their near-term effects and their potential to influence the benefit/cost tests and program designs.

- As benchmarking uses the expertise of engineers who have been working in the field for years, it may be a good approach for many measures, particularly given the large number of measures across all energy efficiency programs. However, past work can be improved upon through the use of more systemized approaches, such as a Delphi-type of analysis.[16]
- Although the billing analyses method addresses the issue of persistence most comprehensively, there are cautions to consider. The effect may be small, which will require large sample sizes. Also, it may be difficult to control for other factors outside the program that cause changes in energy use across a five- or 10-year period. Where quality data exist, a billing analysis is a good method for assessing persistence, but it requires an appropriate data platform for it to be reliable.[17]

3. ***It is important to be open to the new methods and approaches being developed.*** Specifically, a panel of participants established at the time of program participation could be used in cross-sectional, time-series models. This involves incorporating the evaluation of persistence in program design and implementation planning. This type of forward thinking will make persistence easier to address, particularly in near-term years when it is most important.[18]
4. ***Certain types of persistence studies, particularly database/benchmarking approaches, might best be addressed on a regional basis that includes numerous specific programs.*** Assessing persistence across a number of regional programs can provide information on the influence of program design on persistence, which might not be found using a series of program-specific studies. In identifying these regional opportunities, it is important to consider the influence of program design on persistence. (For example, in the study Nexus performed across New England in 2008, program-specific elements had a large influence on the persistence of lighting measures.)

---

[16] Skumatz (2012) presents a number of ways these studies can be improved, including the use of Delphi approaches. An expert-panel approach was used in an evaluation of the Northwest Energy Efficiency Alliance's market transformation programs by Violette and Cooney (Violette and Cooney 2003).
[17] Billing data analyses that try to estimate small effects reliably (for example, 2% savings) without the required sample sizes and accurate data for the independent variables (that is, little measurement error) have often not been successful. Quantum (Quantum 1998) discusses this issue in the context of using a billing analysis to assess persistence for new home construction.
[18] Panel data methods are suggested as a potential approach in both Skumatz (Skumatz 2012) and Nexus (Nexus 2008).

3 Other Evaluation Issues

This section briefly addresses these evaluation issues: (1) synergy; (2) errors in variables, measurement error, and program tracking; (3) dual baselines; and (4) rebound.

3.1 Addressing Synergies Across Programs

Evaluators are often asked about potential synergies across programs. For example, certain information programs may result in direct savings impacts, but the programs may also be designed to lead participants into other programs. In addition, there may be effects across programs. For example, a whole-house retrofit program may influence the uptake of measures offered in other residential programs. These synergies are useful for designing programs and portfolios. Synergies that increase the overall savings from a portfolio of programs are valuable even if one specific program has lower savings due to these synergies.

The industry practice is to use approximate information to assess the relative importance of synergies. Even this level of analysis has generally been limited in evaluations. However, useful information on synergies can be developed by having evaluators:

1. Identify what they believe may be positive and negative synergies (that is, direction)
2. Determine the rough magnitude of these synergies by benchmarking them as a fraction of the programs' savings.

With this material, portfolio models designed to assess the importance of synergies can produce information useful for assessing investments in energy efficiency and future program/portfolio designs.[19]

[19] This approach does not have to be information intensive in terms of developing useful data for analyzing synergies and benchmarking their magnitude. Two pieces of information are needed: (1) an estimated range of effects, for example, from 5% of program savings to 20% of program savings; and (2) an estimate of where the most likely value falls within this range. Based on these three points—the lower bound, the upper bound, and an estimate of where within this range the most likely value falls—Monte Carlo methods can be used to test the importance and sensitivity of program impacts to identified synergies using Excel-based tools. An example of this range-based method can be found in Violette and Cooney (Violette and Cooney 2003), and a version of this method is discussed in EPRI (EPRI 2010, p. 5-4). This information can be used by the program administrator to inform the design of future energy efficiency portfolios.

3.1.1 Conclusion

At the present time, the state-of-the practice involves identifying and assessing the potential importance of specific synergies across programs, although this is not always requested of evaluators. If assessing synergies becomes part of an evaluator's reporting requirements, the evaluator could modify surveys to provide useful information on potentially important energy efficiency program design considerations.[20]

3.2 Errors in Variables, Measurement Errors, and Tracking Systems

This section outlines the issues of errors in the input variables to an energy savings calculation. Such errors could be caused by an incorrect engineering calculation or by inaccurate values of the independent variables used in the regression analyses.

It is important that evaluators consider the accuracy of the input data and use the best quality data possible. In this context, data accuracy issues include data that are unbiased on average, but are subject to measurement error. Biased data clearly poses issues for any analysis; however, measurement error in itself poses challenges for evaluation. This is true even when the measurement error may be uncorrelated with the magnitude of the value of the variable, and the error may be equally distributed above and below the true value.

Program implementers need to be aware that the designs of the data tracking system and the data collection processes have a substantial influence on the accuracy and reliability of data. In turn, the accuracy and completeness of the data influence the estimated realization rates and the ability to achieve the target levels of confidence in these estimates.

While errors in variables can bias the evaluation results either up or down, there are several practical factors in energy efficiency evaluations that tend to result in lower realization rates and lower savings estimates. A typical realization rate study uses information from the tracking system to verify that the equipment is in place, working as expected, and achieving the energy savings predicted in the tracking system. Tracking system errors can include not properly recording the site location, contact information, equipment information, location where the equipment is installed, and the operating conditions of the equipment. This will make any associated field verification more difficult and the variance around the realization rate greater.

Different data issues will have different impacts on the estimates; however, improved data quality will usually decrease the variance of the realization rate estimate and increase confidence and precision. When stakeholders have set high target confidence-and-precision levels, it is important to track accurately the essential data (such as the installed measures' location, size, model number, date, contact person) required to produce the initial tracking system estimate of savings at that site.

[20] One reviewer of this chapter pointed out the potential complexities of determining program-specific synergies and their direction "...to the extent that synergies are increasingly observed or acknowledged, policies regarding the use of individual program cost-benefit analysis results for justifying the retention of programs may need to be changed in favor of portfolio level benefit cost analyses." This section was not intended to delve into benefit-cost methods. However, increased attention on synergies across programs is likely to prove useful. Monte-Carlo models that use different scenarios regarding the magnitude and direction of synergies can help assess the robustness of program and portfolio cost-effectiveness.

The issue of errors in variables and measurement error can be important.

- Kennedy (2003) states: "Many economists feel that the greatest drawback to econometrics is the fact that the data with which econometricians work with are so poor."
- Similarly, Chen et al. (Chen et al. 2007) states: "The problem of measurement errors is one of the most fundamental problems in empirical economics. The presence of measurement errors causes biased and inconsistent parameter estimates and leads to erroneous conclusions to various degrees in economic analysis."

Errors in measuring the dependent variable of a regression equation are incorporated in the equation's error term and are not a problem. The issue is with errors in measuring the independent variables used in a regression model. This violates the fixed independent variables assumption of classical linear regression models: the independent variable is now a stochastic variable.[21] A good source for approaches to address the errors-in-variables issue is Chapter 9 in Kennedy (2003).

The program tracking system data used in regression analyses can be a source of potential issues. For example, the inability to track customer participation in multiple programs can cause a number of problems. In these instances, data can be very accurate at the program level, but there is no mechanism to ascertain the effects of participating in multiple programs. For example, if a billing analysis is being conducted of a high-efficiency residential heating, ventilating, and air-conditioning (HVAC) replacement program but the tracking system is not linked to the residential audit and weatherization program that feeds participants into the HVAC program, this will cause bias. When customers first participate in a feeder program but that information is not conveyed in the tracking system used by the HVAC evaluator, then the HVAC program's savings analysis will be biased, most likely on the low side.

Another well-known errors-in-variables issue relates to models that use aggregate data on DSM expenditures and energy consumption in analyzing the relationship between expenditures on energy efficiency activities and changes in energy use.[22] Developing the appropriate datasets poses challenges. For example, Rivers and Jaccard (2011) note that:

> [O]ur data on demand side management expenditures include all demand side management—in particular it includes both load management expenditures as well as energy efficiency expenditures. Since load management expenditures are not aimed at curtailing electricity demand explicitly… (p. 113).

The report then states that they do not believe this is a problem since

> ...utilities that were able to provide us with data (as well as in US utilities), load management expenditures amounted to less than 25% of the total, so error in our estimates should not be too severe, and in particular should not change the nature of our conclusions.

[21] The assumption is that observations of the independent variable can be considered fixed in repeated samples (that is, that it is possible to repeat the sample with the same independent variable values; [Kennedy 2003, p. 49]).
[22] Two recent publications with examples of this are Rivers and Jaccard (Rivers and Jaccard 2011) and Arimura et al. (Arimura et al. 2011).

The authors may be correct, but their assessment was based on judgment with little real analysis of the degree of the issue.

The work by Rivers and Jaccard (Rivers and Jaccard 2011) and by Arimura et al. (Arimura et al. 2011) illustrates the degree of effort often required to develop a useful set of aggregate state/province-level data or utility-level DSM. Using the Energy Information Administration forms, Arimura states: "The original data set has many observations with missing values for DSM spending, even after our meticulous efforts to find them from various sources."[23]

Another issue concerns the fact that numerous states have both utility and third-party program providers, which complicates the development of data that can be used to examine the relationship between utility energy efficiency program expenditures and aggregate energy consumption.

Attenuation bias is a potential issue when there is measurement error in the independent variables used in regression analyses. Simply stated, the implications are these: (1) more noise in the data due to measurement errors will make it more difficult to find significant impacts and (2) those impacts will tend to be biased downwards.[24]

Attenuation bias can be a problem in regression models using independent variables that might have large numbers of measurement errors due to:

- Differences in reporting of values in databases compiled across utilities
- Assignment/allocation of values at a utility service territory level down to a county level to create more observations.

Chen et al. (Chen et. al 2007, 2011) and Satorra (Satorra 2008) present a graphical example of this bias using a measurement error model developed for a simple one-variable regression.

- Using the model $Y = \beta X + e$ and
- having X measured with error,
- the measurement error model $X = x + u$, with x uncorrelated with u, $var(X) = var(x) + var(u)$ can be used to assess the reliability of the estimated coefficient.

The reliability of X is defined as rel = 1 - var(u)/var(X) (which results in a number between 0 and 1).

Satorra performed a set of simulations for a sample size equal to 10 and used different values for the reliability of the regressor X: 1 (accurate), 0.86, 0.61, and 0.50 (considerable measurement error).

[23] See footnotes 15, 16 and 17 in Arimura et al. (2011) for a discussion of the challenges they addressed in developing values of the key variables (that is, the utility's energy efficiency expenditures that could explain changes in energy use and be used to assess cost-effectiveness in terms of cost per kWh saved).

24 This is not a new problem. Chen (2007 and 2011, p. 901) discusses how one of the most famous studies in economics had to address attenuation bias. In his famous book A Theory of the Consumption Function, Milton Friedman (Friedman1957) shows that, because of the attenuation bias, the estimated influence of income on consumption would be underestimated.

Each simulation is shown in Figure 4.

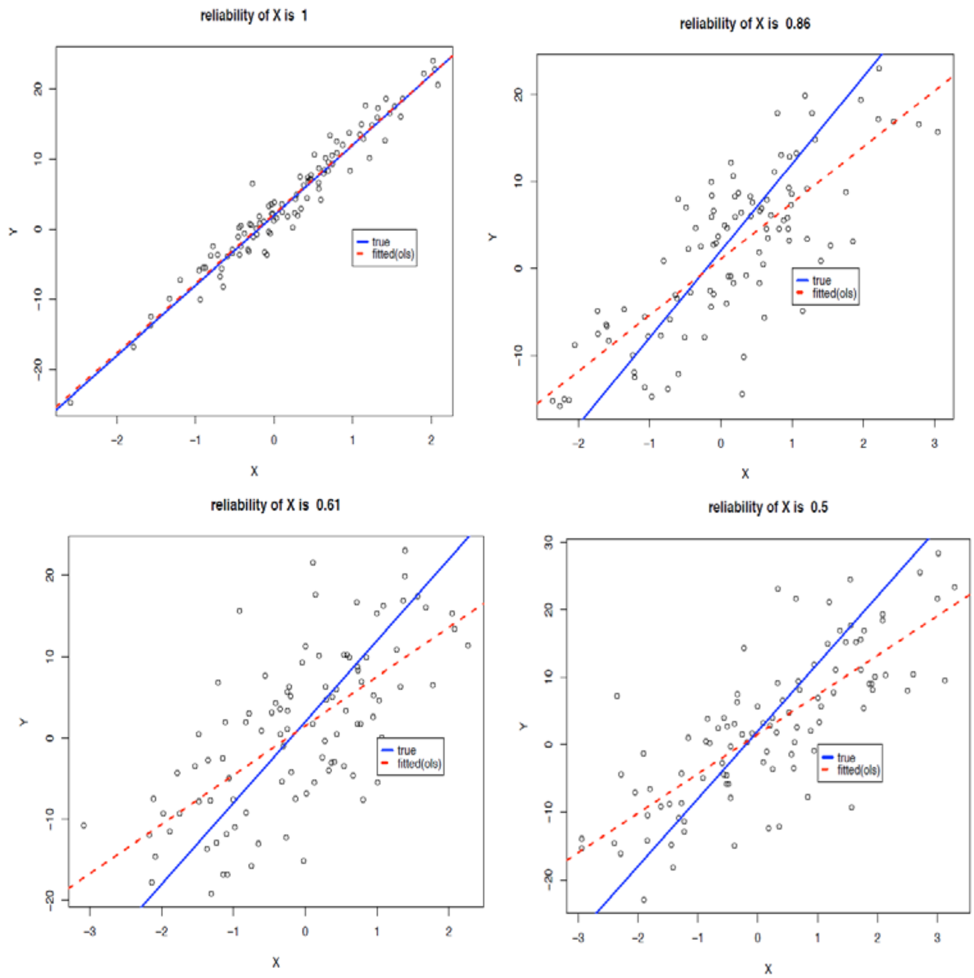


Figure 4. Satorra (2008) Simulation Results

As shown in Figure 4, the bias in the coefficient increases as the reliability of X decreases (that is, measurement error increases), even if this measurement error is uncorrelated with the variance of X. The slope of the coefficient declines as the reliability of X declines. This represents the attenuation bias associated with measurement error.

3.2.1 Conclusion

Issues associated with measurement error are often unavoidable in applied regression analysis. On occasion, data collected for one purpose with one level of accuracy may be used as a variable in a model testing for different types of effects. The solution is to reduce measurement error in the independent variables (the regressors) as much as possible.

Errors in variables, measurement errors, and general issues with data in tracking systems will make it more difficult for the evaluator to identify energy savings at a desired level of confidence. Kennedy (2003) states, “In the spirit of fragility analysis, econometricians should report a range of estimates corresponding to a range of values of measurement variance.”

Kennedy presents examples of how this can be accomplished, but this extra effort is best reserved for large-scale efforts, and it goes beyond current industry standard practice in energy efficiency evaluation.

Nevertheless, having a good data platform from which energy efficiency savings are evaluated is important and needs more emphasis in practical evaluation work.

3.3 Dual Baselines

There are several evaluation issues caused by changes—during the lifetime of that measure—in the baseline against which savings are estimated. One issue, remaining useful life (RUL), occurs when a program is focused on replacing existing (lower-efficiency) equipment with energy efficiency equipment before the old equipment ceases to function *or* before it would otherwise have been replaced. The savings could be:

- Calculated simply as the difference between energy use for the replaced measure and the new energy efficiency measure or
- Based on the difference between the new standard measures available in the market as compared to the new energy efficiency measure.

These savings would be constant for the assumed life of the measure—that is, no adjusted baseline for that measure is considered for the period after the RUL.

In theory, the use of two baselines can be argued to be the appropriate approach in certain applications. The baseline for the replaced low-efficiency measures that still had useful life would be the difference in efficiency between the replaced measure and the high-efficiency measure for the RUL of the replaced measure. For the period *after* the replaced measure's RUL, the baseline should shift to the difference between the installed high-efficiency equipment and the currently available standard equipment. (This would be the baseline for the balance of the assumed life of the new high-efficiency measure.) In practice, this is not often done. (See the conclusions for this section).

A similar situation occurs when a replacement is made of equipment that has a measure life spanning a point when a new code requires higher-efficiency equipment. In this case, evaluators must decide whether the baseline should be the efficiency of the equipment replaced and, in that event, change to a new baseline after the new code or standard is adopted. In general, the working assumption is that the baseline should reflect the energy use of the replaced equipment. If, however, that equipment would have been replaced within a few years by new equipment that meets the new code, then there is a question about whether the baseline should shift.

3.3.1 Conclusions

These dual baseline questions are beginning to receive more attention. Two opinions are expressed in the literature:

- The first and most common is that the complexities and uncertainties entailed in estimating the RULs of the equipment being replaced are excessive compared to their effects on energy savings calculations.
- The second opinion is that dual baseline the issues are important to address for some certain select measures, such as lighting, where the impacts may be large.

These dual-baseline issues have been addressed in some program evaluations, but have not generally been viewed as important for overall energy efficiency program evaluation because of their complexity and uncertainty regarding customer actions. However, the topic of dual baselines deserves more research to assess those specific situations in which accounting for the two baselines might have a substantive effect on energy savings.

3.4 Rebound Effects

Rebound occurs when the costs of using energy are reduced due to energy efficiency programs. When families spend less money to cool their home in the summer because of more efficient equipment, they might change their temperature set point to increase their comfort and their energy use.

Rebound is discussed in the literature according to the following two types:

- ***Type 1: Rebound is used essentially synonymous with take-back*** and happens at the participant level. It involves the question of whether participants who experience lower costs for energy because of an energy efficiency program measure—such as the installation of a high-efficiency air conditioner—then "take back" some of those savings by using more energy.[25]
- ***Type 2: Rebound takes place in the larger economy*** because energy efficiency programs have reduced the cost of energy across a number of uses, stimulating the development and use of energy-using equipment.

With the exception of low-income programs, Type 1 rebound has not been found to be significant in most energy efficiency program evaluations.[26] When consumers match marginal benefits with marginal costs, the concepts of bounded rationality and compartmentalized decision making are being recognized as one theory of consumer behavior and decision

[25] A reviewer pointed out that, for many customers, the lower costs of energy are not reflected in the price of a kWh or a therm of natural gas. Instead, customers use less energy, resulting in a lowering of their monthly bills. This results in customers spending less on energy over the course of a season or year.

[26] This chapter is focused on energy efficiency programs. Take-back is more common in demand response and load management programs where AC units or other equipment are cycled to reduce peak demand for several hours on a few select days. This can result in a warming of the house or building, and the equipment automatically runs a bit more after the cycling event to return the temperature to the original set point. More efficient operational and cycling designs for AC load management programs can greatly reduce take-back, and take-back is a more common effect for event-based load management programs than for energy efficiency programs that influence all hours of a season.

making.[27] (This is contrary to pure economic theory.) Consumers optimize, but only to the point when the complexity of the decision and the cost of the information become too high. For example, although the efficiency of an air-conditioning (AC) unit varies daily with temperature and load; however, a consumer setting the thermostat on the AC unit is probably not going to examine the cost of running that unit each day and then adjust the thermostat accordingly.

Most customers set their thermostats at a comfortable level, regardless of whether they participate in an AC equipment program (whether for maintenance or new equipment) that increases the energy efficiency of the unit. In other words, consumers generally do not change their thermostat setting as a result of participating in an energy efficiency program.

Low-income customers can be the exception, as they may change their thermostat set points for both AC and heating after participating in an energy efficiency program designed to increase the efficiency of the equipment. The change in energy price is more important to low-income customers, who may have been sacrificing comfort to meet their household budget before they participated in the energy efficiency program. Lowering the costs of AC and heating may allow them to set their thermostats at a level that provides more comfort, which may result in greater energy use for this participant segment. While this may cause an increase in the overall energy use for these low-income customers, it can provide a large welfare gain and even improved health and safety for low-income customers.

Going beyond the program participants' actions, Type 2 rebound assesses the economy as a whole, as lowering the cost of energy through aggressive energy efficiency programs may make energy more economical for many new uses. There has been a recent resurgence of interest in this type of rebound, but a full analysis is beyond the scope of this chapter, which focuses on energy efficiency program evaluation. (Gavankar and Geyer [2010] present a review of this larger rebound issue.) There is substantial literature on this economy-wide concept of rebound, and addressing most of the key theses in the discussion requires economy-wide models with energy as one of the inputs for the a wide variety of products and services.[28]

Searching on the terms "energy efficiency" and "rebound" results in many policy papers that present theses on how rebound may be an influence in the larger economy. The issue seems not to be economic welfare, but other policy goals. Using resources as efficiently and cost-effectively as possible always seems like a good policy, unless there is some other constraint. Reducing the cost of energy and allowing people to use energy in additional applications may increase overall welfare. Still, if the goal is to not increase energy use at all, then the downside of reducing energy costs may be concerns about carbon emissions. (It is not the purpose of this chapter, however, to detail this literature, other than noting it exists and offering some practical places to begin a review.)

[27] The primary reference for this concept is Simon (Simon 1957), but it is also discussed in Kahneman (Kahneman 2003).

[28] Other references to discussions of the rebound effect can be found in Vaughn (2012) and in Burns and Potts (2011). Other references are Tierney J. (2011), which presents the issue of rebound as being important, and a counterpoint paper by Afsah (2011.

Using resources as efficiently as possible should be a good start towards any policy designed to reduce energy consumption that may contribute to carbon emissions. This policy could complement pricing and other policies designed to reduce energy use. Starting from a platform of efficient energy use should not hinder the applicability of other policies.

4 References

Allison, P.A. (1995). Survival Analysis Using the SAS System: A Practical Guide. SAS Institute.

Arimura, T.H.; Li, S.; Newell, R.G.; Palmer, K. (2011). Cost-Effectiveness of Electricity Energy Efficiency Programs. National Bureau of Economic Research. NBER Working Paper No. 17556. www.nber.org/papers/w17556.

Afsah, S, K. Salcito and C.Wielga (2011), "Energy Efficiency is for Real, Energy Rebound a Distraction," CO2 Scorecard Research Notes, January. http://co2scorecard.org/Content/uploads/Energy_Efficiency_is_for_Real_CO2_Scorecard_Research_Jan_11_12.pdf

California Public Utilities Commission (CPUC). (2006). California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals. Oregon: TecMarket Works. www.calmac.org/publications/EvaluatorsProtocols_Final_AdoptedviaRuling_06-19-2006.pdf.

Chen, X.; Hong, H.; Nekipelov, D. (2007). "Measurement Error Models." Prepared for the Journal of Economic Literature. www.stanford.edu/~doubleh/eco273B/survey-jan27chenhandenis-07.pdf.

Chen, X.; Hong, H.; Nekipelov, D. (2011). "Nonlinear Models of Measurement Errors." Journal of Economic Literature. (49:4); pp. 901–937. http://cowles.econ.yale.edu/P/cp/p13a/p1344.pdf.

Electric Power Research Institute (EPRI) (January 2010), Methodological Approach for Estimating the Benefits and Costs of Smart Grid Demonstration Projects. Palo Alto, CA: 2010. 1020342. www.smartgridnews.com/artman/uploads/1/1020342EstimateBCSmartGridDemo2010_1_.pdf.

Energy and Resource Solutions (ERS). (November 17, 2005). Measure Life Study. Prepared for The Massachusetts Joint Utilities.

Friedman, M. (1957). A Theory of the Consumption Function. Princeton University Press. www.nwcouncil.org/energy/rtf/subcommittees/comlighting/Measure%20Life%20Study_MA%20Joint%20Utilities_2005_ERS-1.pdf.

GDS Associates (2007). Measure Life Report: Residential and Commercial/Industrial Lighting and HVAC Measures. Prepared for The New England State Program Working Group for use as an Energy Efficiency Measures/Programs Reference Document for the ISO Forward Capacity Market. http://neep.org/uploads/EMV%20Forum/EMV%20Studies/measure_life_GDS%5B1%5D.pdf.

Gavankar, S.; Geyer, R. (June 2010). The Rebound Effect: State of the Debate and Implications for Energy Efficiency Research. Institute of Energy Efficiency (UCSB). http://iee.ucsb.edu/files/pdf/Rebound%20Report%20for%20IEE-UCSB_0.pdf.

Kahneman, D. (2003). “Maps of Bounded Rationality: Psychology for Behavioral Economics.” *The American Economic Review.* (93:5); pp. 1449–75. www.econ.tuwien.ac.at/lotto/papers/Kahneman2.pdf.

KEMA. (2004). Residential Refrigerator Recycling Ninth Year Retention Study. Study ID Nos. 546B, 563, Madison. www.calmac.org/results.asp?t=2.

KEMA. (August 25, 2009). Focus on Energy Evaluation: Business Programs: Measure Life Study Final Report. Prepared for the Public Service Commission of Wisconsin. www.focusonenergy.com/files/Document_Management_System/Evaluation/bpmeasurelifestudyfinal_evaluationreport.pdf.

Kennedy, P. (2003). A Guide to Econometrics. Chapter 9. MIT Press.Navigant Consulting, Inc. (February 20, 2011). Evaluation Report: OPower SMUD Pilot Year 2. Presented to OPower. http://opower.com/uploads/library/file/6/opower_smud_yr2_eval_report_-_final-1.pdf.

Nexus Market Research, Inc. (June 4, 2008). Residential Lighting Measure Life Study. Cambridge: New England Residential Lighting Program. www.puc.nh.gov/Electric/Monitoring%20and%20Evaluation%20Reports/National%20Grid/121_NMR_Res%20ltg%20measure%20life.pdf.

Peterson, G.; deKieffer, R.; Proctor, J.; Downey, T. (1999). Persistence #3A: An Assessment of Technical Degradation Factors: Commercial Air Conditioners and Energy Management Systems, Final Report. CADMAC Report # 2028P.

Burns, C; Potts, M (2011). The “Rebound Effect”: A Perennial Controversy Rises Again. Solutions Journal: Spring 2011 (Vol. 4, No. 2) http://www.rmi.org/TheReboundEffectAPerennialControversyRisesAgain.

Vaugh, K. (2012). Jevons Paradox: The Debate That Just Won't Die – Response by Amory Lovins, RMI Outlet, March 20. http://blog.rmi.org/blog_Jevons_Paradox

Proctor Engineering. (1999). Summary Report of Persistence Studies: Assessments of Technical Degradation Factors, Final Report. CADMAC Report #2030P. www.calmac.org/publications/19990223CAD0003MR%2EPDF.

Quantum Consulting. (1998). PG&E Statewide Multi-Year Billing Analysis Study: Commercial Lighting Technologies Final Report. Berkeley.

Rivers, N.; Jaccard, M. (2011). “Electric Utility Demand Side Management in Canada.” The Energy Journal. (32:4); pp. 93-116. http://bwl.univie.ac.at/fileadmin/user_upload/lehrstuhl_ind_en_uw/lehre/ws1112/SE_Int._Energy_Mgmt_1/DSMCanada.pdf.

Satorra, A. (2008). Theory and Practice of Structural Equation Modeling, Department d'Economiai Empresa. Barcelona. http://statmath.wu.ac.at/courses/TPStrucEqMod/errorinvariables.pdf.

Simon, H. (1957). A Behavioral Model of Rational Choice in Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting. New York: Wiley.

Skumatz, L.A. et al. (2009). Lessons Learned and Next Steps in Energy Efficiency Measurement and Attribution: Energy Savings, Net to Gross, Non-Energy Benefits, and Persistence of Energy Efficiency Behavior. http://uc-ciee.org/downloads/EEM_A.pdf.

Skumatz, L.A. (2012). What Makes a Good EUL? Analysis of Existing Estimates and Implications for New Protocols for Estimated Useful Lifetimes (EULs). International Energy Program Evaluation Conference. Rome, Italy.

Tierney J. (2011), "When Energy Efficiency Sullies the Environment", The New York Times, March 2011.

Violette, D.; Cooney, K. (December 8, 2003). Findings and Report: Retrospective Assessment Of The Northwest Energy Efficiency Alliance. Prepared for Northwest Energy Efficiency Alliance Ad Hoc Retrospective Committee. Summit Blue/Navigant Consulting. www.theboc.info/pdf/Eval-BOC_SummittBlue_NEEA_2003.pdf.

5 Resources

Ahmad, M.; Deng, A.; Spoor, S.; Usabiaga, M.; Zhao, I. (2011). *Persistence of Energy Savings in Industrial Retrocommissioning Projects*. ACEEE Summer Study on Energy Efficiency in Industry. www.aceee.org/files/proceedings/2011/data/papers/0085-000028.pdf.

Decision Sciences Research Associates. (1999). *1994 Commercial CFL Manufacturers' Rebate Persistence Study.* Report ID 529D. Pasadena: Decision Sciences Research Associates.

KEMA. (April 24, 2006). *2005 Smart Thermostat Program Impact Evaluation*. Final report prepared for San Diego Gas and Electric Company. http://sites.energetics.com/madri/toolbox/pdfs/pricing/kema_2006_sdge_smart_thermostat.pdf.

Navigant Consulting, Inc. (2010). *Kaizen Blitz Pilot. Report 1.* Prepared for Energy Trust of Oregon.

Ontario Power Authority. (2011). *EM&V Protocols and Requirement*. Ontario: 2011-2014.

Pacific Gas & Electric. (1999). *Commercial Lighting Study.* CALMAC. pp. 4-14.

Proctor Engineering. (1998). *Statewide Measure Performance study #2: An Assessment of Relative Technical Degradation Rates. Final Report.* For California Measurement Advisory Committee.

RLW Analytics. (1998). *SCE Non-Residential New Construction Persistence Study. Final Report.* For Southern California Edison. Study # SCE0064.01; 554; 530.

San Diego Gas & Electric. (March 2006). *1996 & 1997 Commercial Energy Efficiency Incentives Ninth Year Retention Evaluation.* www.calmac.org/publications/2006%5FPY96%26PY97%5FCEEI%5F9th%5FYear%5FRetention%5FEvaluation%2Epdf.

U.S. Environmental Protection Agency. (November 2007). *Model Energy Efficiency Program Impact Evaluation Guide.* www.epa.gov/cleanenergy/documents/suca/evaluation_guide.pdf.

6 Appendix A: Program-Specific Persistence Study Challenges and Issues[29]

Persistence studies provide useful information for making sensible energy efficiency (EE) investment decisions when the benefit/cost test of a measure is sensitive to changes in savings over time. As such, various persistence study challenges and issues should be examined regarding how energy savings are estimated (e.g., through measure and/or behavioral change). Table 3 summarizes persistence study challenges and issues by energy activity.

Table 3. Persistence Study Challenges and Issues

Program Measure or Activity	Characteristics	Persistence Study Challenges and Issues
New Installation, Retrofit, and Replace on Burnout	• Intervention occurs at the time measures are being replaced. • Savings result from the difference in energy use between the old equipment and the EE equipment. • An example is a lighting rebate program that provides incentives to participants for switching to higher-efficiency lighting measures.	• Cost of on-site data collection is high. • Impractical to wait for half of the units to fail so as to determine median survival time. • Some owners prematurely interrupt measure life for various reasons (such as dissatisfaction with new equipment) and switch back to less-efficient equipment. • Measure life estimates are based on failures. However, as there are few equipment failures in the early stages of equipment life, it is difficult to get an unbiased determination of expected useful life (EUL). • A lack of plug load sector data. • Business turnover has a strong effect on commercial measure lifetime. • When replacing equipment before the end of equipment life, the question of whether EE should be calculated by the delta of efficient equipment compared either to (1) replaced equipment, or (2) the equipment required by codes and standards. There is difficulty in predicting future standards.
Early Retirement	• Accelerates the retirement of inefficient equipment. • Savings result from load reduction due to absence of inefficient equipment. • An example is a refrigerator recycling program that gives participants an incentive for terminating the use of inefficient refrigerators.	• RUL is not well-studied, thus, it introduces uncertainties to future savings after the early retirement of the old equipment.

[29] Ms. Angie Lee and Mr. Mohit Singh-Chhabra of Navigant, Inc., developed this appendix.

Behavioral Programs		
Energy Activity	**Characteristics**	**Current Persistence Study Challenges and Issues**
Feedback[30]	• Programs that influence behavioral changes to obtain energy savings. • Savings result from behavioral changes. • An example is an informational program that tells households of their energy consumption as compared to their neighbors.	• Current standard behavior is going to change, and future standard behavior is difficult to predict. • A lack of studies on behavioral programs. • It is difficult to find an unbiased, uncontaminated control group.
Educational/Training	• Educational programs that provide customers with EE education. • Savings result from behavioral changes. • An example is a school education program.	• Current standard behavior is going to change, and future standard behavior is difficult to predict. • A lack of studies on behavioral programs.
Operation & Maintenance (O&M)	• Provides O&M best practices with low-cost/no-cost measures, such as adjusting control settings. • Savings result from improved O&M. • An example is retro-commissioning activity.	• Retro-commissioning programs typically have a short useful life[31], since most of the activities involve adjusting controls. • Operators who are unaware of the reason behind adjustments could revert back to the original settings.

[30] Navigant Consulting (2011).
[31] Ahmad et al. (2011).

Table 4. Measure and Behavioral Programs

Measure and Behavioral Programs		
Energy Activity	**Characteristics**	**Current Persistence Study Challenges and Issues**
Whole Building New Construction and Retrofit[32]	• Combination of both EE measures and O&M best practices. • Savings result from the difference in energy use between the old equipment and the EE equipment, as well as from O&M best practices over baseline behavior.	• It is difficult to separate out the effects of specific measures in a whole-building system, as most energy evaluations utilize billing analysis or building simulations to estimate whole-building savings.
Smart Thermostat[33]	• Thermostats are used to influence AC use. • Users obtain incentives for allowing the utility to adjust their thermostat set points while reserving the right to override the utility re-set. • Savings result from reduction in energy usage occurring from changes in AC use.	• A lack of persistence studies on smart thermostat programs.

[32] RLW Analytics (1998).
[33] KEMA (2006).

The following table presents candidate methods by study type—measure life, retention and degradation.

Table 5. Methodology Summary

Method	Method Description and Application	Data Requirements	Applicable Studies		
			Measure Life	Retention	Degradation
On-Site Equipment Installation Verification	• Verifications through an on-site inspection: (1) that equipment is in-place and operable, and (2) whether the application of the equipment has changed. • Applicable to evaluating measure programs. • An example is a measure life/EUL study of a commercial lighting incentive program using on-site audits[34].	• Measure make and model. • Spot measurements to supplement visual inspection. • Date installed and date when measure became inoperable or was removed.	x	x	
On-site Equipment Measurement and Testing	• Measurement (short term or long term) of equipment performance, focused on collecting data and ensuring equipment is use as designed. If it is not, then identifying the reasons the usage differs from the equipment's design intent.) • Applicable to evaluating measure programs. • An example is a degradation study of high-efficiency motors.	• Measure make and model. • Use of equipment as designed. • Observation of failure rates.			x
Laboratory Testing	• Measurement of energy use of both EE and standard equipment over time in unoccupied facilities. • Laboratory testing must account for the operational conditions expected for installations. • Applicable to evaluating measure programs. • An example is a degradation study comparing existing and high-efficient air compressors.	Energy use of equipment over time.			x

[34] San Diego Gas & Electric (1999).

Benchmarking and Secondary Literature Review	• Engineering review of equipment degradation and uncertainties. The literature search should include journal articles, conference proceedings, manufacturer publications, and publications of engineering societies. • Applicable to evaluating both measure and behavioral programs. • An example is an assessment of measure technical degradation rates by conducting a meta-review on secondary literature.[35]	Equipment and/or behavior degradation and uncertainties.	x	x	x
Telephone Surveys/ Interviews	• Interviews of program participants about: (1) their consumption patterns compared to EE equipments' design intent, and (2) whether the EE equipment is in place and operable. • Applicable to evaluating both measure and behavioral programs. • An example is a persistent study of an O&M program studying behavioral retention.[36]	Equipment failures and/or replacement behavior, including time of failure and/or replacement, and the number of failures and/or replacements.	x	x	x
Billing Analyses – Fixed Effects and Statistically Adjusted Engineering Models[37]	• Statistical analysis to model the difference between customers' energy usage pre- and post-analysis periods, using real customer billing data over multiple years. • Applicable to measure and behavioral programs. • An example is evaluating multiyear savings persistence on commercial lighting technologies.[38]	Customer billing data over time.		x	x

[35] Proctor Engineering (1998).
[36] Navigant Consulting, Inc. (2010).
[37] Pacific Gas & Electric (1999).
[38] Quantum Consulting (1998).

Survival Curves	• Linear, logistics, exponential, or hazard models estimating equipment survival rate. The model choice depends on equipment characteristics and previous research. • Applicable to measure and behavioral programs. • An example is estimating the EUL of equipment installed in a new construction project using survivor function and hazard function.	Independence of equipment failure and EUL.	x		
Controlled Experiment	• Experiment developed across census, randomly assigning participants into treatment and control groups. • Applicable to behavioral programs. • An example is a retention study of a behavioral energy program over multiple years.	Customer billing data of control group and treatment group over time.		x	x

Chapter 14: Chiller Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in September 2014.

Alex Tiessen
Posterity Group
Ottawa, Ontario

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68570
September 2017

Contract No. DE-AC36-08GO28308

Chapter 14: Chiller Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in September 2014.

Alex Tiessen
Posterity Group
Ottawa, Ontario

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68570
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter author wishes to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Greg McGuire of ICF International.

Suggested Citation

Tiessen, A. (2017). *Chapter 14: Chiller Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68570. http://www.nrel.gov/docs/fy17osti/68570.pdf

Acronyms

ANSI	American National Standards Institute
BAS	building automation system
EUL	effective useful life
HVAC	heating, ventilating, and air conditioning
IPLV	integrated part load value
IPMVP	International Performance Measurement and Verification Protocol
kWh	kilowatt-hour
M&V	measurement and verification
OAT	outdoor air temperature
RUL	remaining useful life
TMY	typical meteorological year
TRM	technical reference manual

Table of Contents

List of Figures

List of Tables

1 Measure Description

This protocol defines a chiller measure as a project that directly impacts equipment within the boundary of a chiller plant. A chiller plant encompasses a chiller—or multiple chillers—and associated auxiliary equipment. This protocol primarily covers electric-driven chillers and chiller plants. It does not include thermal energy storage and absorption chillers fired by natural gas or steam, although a similar methodology may be applicable to these chilled water system components.[1]

Chillers provide mechanical cooling for commercial, institutional, multiunit residential, and industrial facilities. Cooling may be required for facility heating, ventilation, and air conditioning (HVAC) systems or for process cooling loads (e.g., data centers, manufacturing process cooling).

The vapor compression cycle,[2] or refrigeration cycle, cools water in the chilled water loop by absorbing heat and rejecting it to either a condensing water loop (water cooled chillers) or to the ambient air (air-cooled chillers). As listed in Table 1, ASHRAE standards and guidelines define the most common types of chillers by the compressors they use (ASHRAE 2012).

Table 1. Four Common Chiller Types

Chiller Type	Description
Reciprocating, Screw, and Scroll	Reciprocating, screw, and scroll chillers use positive-displacement compressors. These compressors increase refrigerant vapor pressure by reducing the volume of the compression chamber. Reciprocating chillers compress air using pistons; screw chillers compress air using either single- or twin-screw rotors with helical grooves; and scroll chillers compress air through the relative orbital motion of two interfitting, spiral-shaped scroll members.
Centrifugal	Centrifugal chillers use dynamic compressors. These compressors increase refrigerant vapor pressure through a continuous transfer of kinetic energy from the rotating member to the vapor, followed by the conversion of this energy into a pressure rise. Centrifugal chillers transfer this kinetic energy using impellers similar to turbine blades.

Chiller plant auxiliary equipment includes chilled water and condensing water pumps; cooling tower fans and spray pumps (water-cooled chillers); condenser fans (air-cooled chillers), and water treatment systems.

Projects impacting chiller plant equipment generally fall into one of two categories:

- **Equipment replacement.** These projects involve replacing a chiller and possibly replacing some or all of the auxiliary equipment.

[1] As discussed in the section "Considering Resource Constraints" of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.
[2] The vapor compression cycle consists of four main components: an evaporator, a compressor, a condenser, and an expansion valve.

- **Modifications to existing equipment.** These projects typically involve adding control equipment (e.g., adding a variable frequency drive to an existing centrifugal chiller to improve its part-load efficiency).

2 Application Conditions of Protocol

A program may address chiller energy-efficiency activities alone, but more often, broader commercial, multiunit residential, or industrial custom programs will include these activities. As chiller savings often occur at the same time many jurisdictions experience electricity system peaks, savings from these projects can have a significant impact on a custom program's summer peak-demand savings.

Service providers and other stakeholders design energy-efficiency programs to overcome market barriers through activities that address the available market opportunities. Chiller programs may include some or all of the following activities:

- **Training.** Program administrators sometimes fund or develop training for service providers. For example, in some jurisdictions, service providers do not routinely undertake detailed common practice, feasibility studies for their customer base. If a program is to exploit to the fullest extent the achievable potential in its region, end users need to consider early replacement of equipment in their chiller plants. To facilitate this decision-making process, service providers may need training on how to conduct investment-grade energy audits, using recommended practices.
- **Development incentives.** Program administrators sometimes provide incentives that encourage end users to undertake detailed feasibility studies for chiller measures. Ideally, the incentives encourage end users to commission a detailed feasibility study, which could result in the development of a business case that would encourage end users to move forward with a chiller measure.
- **Implementation incentives.** Program administrators often provide incentives to implement chiller measures. Again, ideally, the incentives can encourage end users to invest more capital upfront to install higher-efficiency equipment or to invest capital sooner in early replacement projects.

This protocol provides direction on how to reliably verify savings from chiller measures using a consistent approach. It does not address savings achieved through training or through market transformation activities.

3 Savings Calculations

This section presents a high-level gross energy savings equation[3] that applies to all chiller measures. Section 4, *Measurement and Verification Plan,* provides detailed direction on how to apply this equation.

Use the following general equation to determine savings (US DOE FEMP 2008).

Equation 1

$$kWh\ Savings_{Total} = (kWh\ Savings_{Chiller}) + (kWh\ Savings_{Auxiliary})$$

Where,

$kWh\ Savings_{Total}$ = First-year energy consumption savings

$kWh\ Savings_{Chiller/Auxiliary}$ = $\sum_{Cooling\ Load\ Range}\left(kWh_{Baseline} - kWh_{Reporting}\right)_{Cooling\ Load}$

And,

$kWh_{Baseline,\ Cooling\ Load}$ = Energy required by the baseline equipment (either existing or hypothetical) at a given cooling load

$kWh_{Reporting,\ Cooling\ Load}$ = Energy required by the new equipment at a given cooling load

The approach for determining demand savings for chiller measures depends on the type of load being served by the chiller plant:

- **HVAC loads.** For chillers serving HVAC loads, apply regional load savings profiles based on regional weather (average daily load profiles for each season), calibrated building simulation models, engineering models targeting peak demand periods, and/or peak coincident factors to consumption savings data.
- **Process loads.** As load savings profiles vary, depending on the process, calculating the demand savings for chillers serving process loads is not as straightforward as it is for chillers serving HVAC loads. First, produce project-specific load savings profiles and then apply site-specific coincidence factors to determine coincident peak demand savings.

3.1 Determining Baseline Consumption

A common issue for many chiller programs is the use of existing equipment in determining the baseline for establishing project savings claims. The following discussion explains why this is not always the correct baseline.

[3] As presented in the Introduction, the protocols focus on gross energy savings and do not include other parameter assessments, such as net-to-gross, peak coincidence factors, or cost-effectiveness.

To establish an appropriate baseline, consider three main replacement scenarios (Fagan et al. 2011):

- **Early replacement.** Existing equipment has a remaining useful life (RUL).
- **Replace-on-burnout.** The effective useful life (EUL) of the existing equipment has expired.
- **Natural turnover.** Replacement of equipment for reasons other than energy savings.

For the first scenario (early replacement), apply a dual baseline (Ridge et al. 2011), as shown in Figure 1. For the latter two scenarios, establish a hypothetical baseline that uses a new chiller meeting the applicable energy-efficiency standard[4] for the applicable jurisdiction. The hypothetical baseline should also consider industry standard practices and the existing equipment, which may set higher efficiency levels than the applicable energy-efficiency standards.

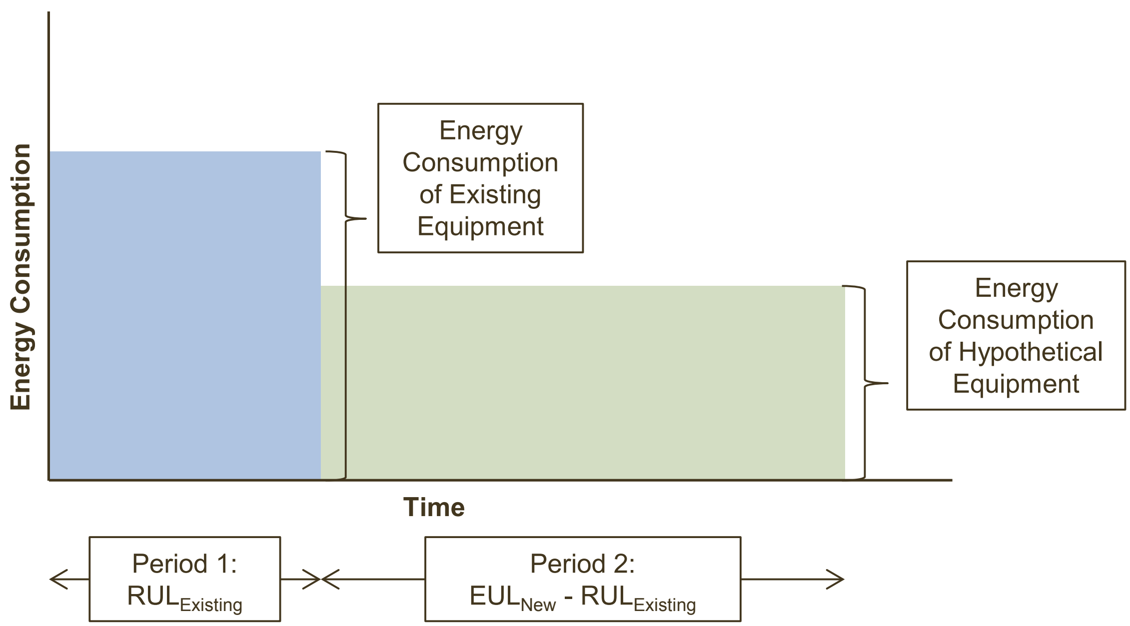


Figure 1. Dual baseline

As shown in Figure 1, there are two distinct baseline periods:

- **Period 1.** For the duration of the RUL of existing equipment, the existing equipment is the baseline.
- **Period 2.** For the remaining EUL of new equipment, use a hypothetical baseline.

As available, use the program defined EUL for chiller equipment or consult regional technical reference manuals (TRM); when program or TRM information is not available, use other

[4] American National Standards Institute (ANSI)/ASHRAE Standard 90.1 is an example of a widely recognized energy-efficiency standard.

secondary sources.[5] Similarly, use the method defined by the program to determine the RUL of baseline chiller equipment. If this has not been previously established, consider defining RUL as the difference between the EUL and current age of the chiller (or number of years since its last rebuild)[6].

[5] California's Database for Energy Efficient Resources suggests an EUL of 20 years for chillers (CPUC 2008).
[6] Evaluators should use discretion regarding the scope of the rebuild and how it may impact the RUL of the chiller.

4 Measurement and Verification Plan

This section contains both recommended approaches to determining chiller energy savings and the directions on how to use the approaches under the following headings:

- Measurement and verification (M&V) method
- Data collection
- Interactive effects
- Detailed procedures
- Regression model direction.

4.1 Measurement and Verification Method

This protocol recommends an approach for verifying chiller energy savings that adheres to Option A of the International Performance Measurement and Verification Protocol (IPMVP). Because it is not possible to measure performance data for hypothetical baseline equipment, this protocol recommends Option A (retrofit isolation—key parameter measurement) rather than Option B (retrofit isolation—all parameter measurement).

Key parameters that require measurement include cooling load data and independent variable data, such as outdoor air temperature (OAT). Estimated parameters include manufacturer part-load efficiency data.[7]

In some cases, metered data may be available directly from the facility's building automation system (BAS).[8] Also, if required, the facility can add control points to the BAS, either as part of the implementation process or specifically for M&V purposes. Where the BAS cannot provide information, the protocol recommends using submeters and data loggers to collect data.

To ensure the M&V method balances the need for accurate energy savings estimates with the need to keep costs in check (relative to project costs and anticipated energy savings), consider two alternate approaches—IPMVP's Option C and Option D.

- **Option C.** Consider a whole-facility approach for early replacement projects if metering the required parameters is cost-prohibitive *and* if the estimated project-level savings are large compared to the random or unexplained energy variations that occur at the whole-facility level.[9] This approach is relatively inexpensive because it involves an analysis of facility consumption data. The downside is evaluators cannot perform verification until after collecting a full season or year of reporting period data and monitoring and

[7] Even though evaluators can measure efficiency data for the reporting period, under a hypothetical baseline scenario it is generally recommended to use pre- and postinstallation manufacturer efficiency data. This approach provides a more accurate estimate of the change in efficiency in comparison to an approach that uses a combination of measured reporting period efficiency data and manufacturer baseline efficiency data.

[8] It is important to ensure qualified service personnel maintain the BAS. Transducers that are out of calibration, or simply broken, could significantly impact M&V results.

[9] Typically, savings should exceed 10% of the baseline energy for the facility's electricity meter to confidently discriminate the savings from the baseline data when the reporting period is shorter than two years (EVO 2012).

documenting any changes to the facility's static factors[10] over the course of the measurement period. Also, an analysis of monthly consumption data may be inadequate for estimating peak demand savings; evaluators should investigate whether data from advanced metering infrastructure (e.g., interval meters) is available to increase the accuracy of billing data analyses.

- **Option D.** Consider a calibrated simulation approach if metering the required parameters is cost-prohibitive *and* the estimated project-level savings are small compared to the random or unexplained energy variations that occur at the whole-facility level. Undertake calibration in two ways: (1) calibrate the simulation to actual baseline or reporting period consumption data and (2) confirm the reporting period inputs via the BAS front-end system or the chiller control terminal, when possible. [11,12]

4.2 Data Collection

When using Option A (the preferred approach) to assess chiller measures, the following M&V elements require particular consideration:

- Measurement boundary
- Measurement period and frequency
- Functionality of the measurement equipment
- Savings uncertainty.

4.2.1 Measurement Boundary

For all projects, especially those that require metering external to the BAS, it is important to define the measurement boundary. When determining boundaries, consider the location and number of measurement points required as well as the project's complexity and expected savings:

- A narrow boundary simplifies data measurement (e.g., chiller plant equipment directly affected by the chiller measure), but will require accounting for any variables driving energy use outside the boundary (interactive effects)[13]
- A wide boundary will minimize interactive effects and increase accuracy. However, since M&V costs may also increase, it is important to ensure the expected increase in the accuracy of the project savings justifies the M&V cost increase.

---

[10] Many factors can affect a facility's energy consumption even though evaluators do not expect them to change. These factors are known as "static factors" and include the complete collection of facility parameters that are generally expected to remain constant between the baseline and reporting periods. Examples include: building-envelope insulation, space use within a facility, and facility square footage.
[11] In many cases, the simulation should represent the entire facility; however, in some cases, depending on the facility's wiring structure, evaluators can apply a similar approach to building submeters, such as distribution panels that include the affected systems.
[12] See the Uniform Methods Project's *Commercial New Construction Protocol* for more information on using Option D.
[13] Although significant interactive effects are uncommon for chiller measures, there are some scenarios that warrant consideration. See Section 4.3 for further detail.

4.2.2 Measurement Period and Frequency

Consider these important timing metrics: (1) the measurement period and (2) the measurement frequency. In general:

- Choose the measurement period (the length of the baseline and reporting periods) to capture a full cycle of each operating mode. For example, if a chiller is serving an HVAC load, collect data over the summer, shoulder, and winter seasons (if applicable).
- Choose the measurement frequency (the regularity of measurements during the measurement period) by assessing the type of load:
  - **Spot measurement.** For constant loads (e.g., constant-speed chilled water pumps), measure power briefly, preferably over two or more intervals.
  - **Short-term measurement.** For loads predictably influenced by independent variables (e.g., chiller compressors serving HVAC loads), take short-term consumption measurements over the fullest range of possible independent variable conditions, given M&V project cost and time limitations.
  - **Continuous measurement.** For variable loads (e.g., chiller compressors serving process loads), measure consumption data continuously, or at appropriate discrete intervals, over the entire measurement period.

Section 4.4, *Detailed Procedures*, provides directions regarding measurement period and frequency for each element of the previously introduced savings equation.

4.2.3 Measurement Equipment

When the BAS cannot provide enough information and submeters are necessary to obtain data, use these guidelines to select the appropriate meter:[14]

- Size the meter for the range of values expected most of the time.
- Select the meter repeatability and accuracy that fits the budget and intended use of the data.
- Install the meter as recommended by the manufacturer.
- Calibrate the meter before it goes into the field and maintain meter calibration, as recommended by the manufacturer. If possible, select a meter with a recommended calibration interval that is longer than the anticipated measurement period.
- If budget allows, consider installing submeters permanently.

If using BAS data, exercise due diligence by determining when the BAS was last calibrated and by checking the accuracy of the BAS measurement points.

Table 2 lists recommended levels of accuracy for the types of metering equipment used for chiller M&V (US DOE FEMP 2008).

[14] Further information on choosing meters can be found in the Uniform Methods Project's *Metering Cross-Cutting Protocols.*

Table 2. Recommended Meter Accuracies

Meter Type	Purpose	Accuracy of Meter
Flow meter	Chilled water flow (GPM)	± 2%
Immersion temperature sensors	Chilled water temperatures	± 0.3°F
Power meters	True RMS power (kW)	± 2%
Outdoor air temperature sensors	Outdoor air dry bulb temperatures	±1.0°F

4.2.4 Savings Uncertainty

If possible, quantify the accuracy of measured data[15] and, if practical, conduct an error propagation analyses to determine overall impacts on the savings estimate.

4.3 Interactive Effects

For projects evaluated using Option A, consider and estimate any significant interactive effects. Although significant interactive effects are uncommon for chiller measures, there are some scenarios that warrant consideration. For example, if a facility uses waste heat from a chiller plant (heat taken from the condenser loop) to satisfy coincident heating loads, then a chiller measure that increases the efficiency of the chiller plant will decrease the amount of waste heat available. In such cases, estimate interactive effects by using equations that apply the appropriate engineering principles.

Interactive effects for projects being verified using Option C or Option D are typically included in the facility-level savings estimates.

4.4 Detailed Procedures

This section lists the detailed steps required for using the recommended M&V approach (Option A) for chiller measures (specifically, for projects that impact both chillers and the chiller's auxiliary equipment).

4.4.1 Chillers

Table 3 presents the five-step procedure for determining the chiller savings term in Equation 1 ($\text{kWh Savings}_{\text{Total}} = \text{kWh Savings}_{\text{Chiller}} + \text{kWh Savings}_{\text{Auxiliary}}$). These steps cover the range of actions depending on:

- Whether the chiller plant is serving an HVAC load or a process load or
- Whether the plant has a single schedule or multiple operating schedules.

[15] Metering accuracy is only one element of savings uncertainty. Inaccuracies also result from modeling, sampling, interactive effects, estimated parameters, data loss, and measurements being taken outside of a meter's intended range.

Table 3. Chiller M&V Procedures

<table>
<tr><th>Step</th><th>Details</th></tr>
<tr><td>Develop load curve model(s) by measuring reporting period operation</td><td>
To calculate chilled water load, use coincident measurements of chilled water flow (gpm), and chilled water supply and return temperatures ($^{\circ}F$):
Cooling load (tons) = 500(gpm)(ΔT $^{\circ}F$)/(12,000 BTUh/ton)
<u>For HVAC loads</u>: Take (or collect) short-term measurements at representative load levels for each season (summer, shoulder, winter) and for each schedule type, if applicable. Evaluator may also collect chilled water flow and chilled water temperatures by the BAS and calculated cooling load (BTUh or tons) directly by the BAS.
<u>For process loads</u>: Take continuous measurements over the length of each type of process cycle.
Additionally, collect the independent variable data:
<u>For HVAC loads</u>: Measure or collect coincident site-specific OAT dry-bulb and wet-bulb data.
<u>For process loads</u>: Measure or collect coincident process data.[a]
Conduct a regression analysis to determine the relationship between independent variables and cooling load—this relationship should be expressed in terms of an equation (load curve model). Evaluators may be required to run multiple regression models. For example, if the chiller plant is serving an HVAC load and has an occupied and an unoccupied schedule (e.g., an occupied cooling set point temperature, and an unoccupied cooling set point temperature), evaluators may require two regression models.
</td></tr>
<tr><td>
<u>For HVAC Loads</u>:
Develop a bin operating profile by typical meteorological year (TMY)[b] OAT data or, if possible, develop an hourly profile over the full operating schedule of the affected equipment
<u>For Process Loads</u>:
Develop a bin operating profile by normalized process data
</td><td>
If a bin analysis is being used, develop bin data tables that present the following data (one table for each schedule type, if applicable):
HVAC Load
<table>
<tr><th>Independent Variable</th><th>Load</th><th>Annual Hours</th></tr>
<tr><td>Create approximately 10 OAT bins over the TMY data range</td><td>Calculate the normalized load by applying the load curve model to the midpoint of each temperature bin</td><td>Base this on TMY data and the chiller operating schedule</td></tr>
</table>
Process Load
<table>
<tr><th>Independent Variable</th><th>Load</th><th>Annual Hours</th></tr>
<tr><td>Create an appropriate number of process level bins for the given process parameter range</td><td>Calculate the normalized load by applying the load curve model to the midpoint of each bin</td><td>Use continuous measured data to estimate the hours of operation within each bin</td></tr>
</table>
If an hourly analysis is being used for HVAC loads, the normalized load for each hour should be calculated by applying the load curve model developed in Step 1. In this scenario, the subsequent analysis outlined in Steps 3 through 5 should be conducted on an hourly basis, rather than on a bin-by-bin basis.
</td></tr>
</table>

Step	Details
Apply manufacturer part-load efficiency data to the bin data	Apply kilowatt/ton part-load efficiency data from manufacturer specification sheets to each bin and then calculate kilowatt-hour as follows: $kWh_{bin} = tons_{bin} \times hrs_{bin} \times kW/ton_{bin}$ Do this for the baseline (both existing and hypothetical if a dual baseline is applicable) and the new chiller *for each schedule type, if applicable.* The part-load efficiency data presented by manufacturers is typically calculated based on Air-Conditioning and Refrigeration Institute standard conditions. If available, use manufacturer efficiency data that adjusts for designer-specified evaporator and condenser entering and leaving water temperatures. *If part-load efficiency data does not align with bin mid-points, interpolate. *If part-load efficiency data does not exist for the baseline chiller, apply the integrated part load value (IPLV) to all bins.
Calculate kilowatt-hour savings for each bin for each schedule type	For each schedule type: $kWh\ Savings_{bin} = kWh_{bin,\ Baseline} - kWh_{bin,\ Reporting\ Period}$
Sum kilowatt-hour savings across all load bins for each schedule type	For each schedule type: $$\sum_{Bin\ Data\ (Coolinlg\ Load) Range} kWh\ Savings_{Bin\ (Cooling\ Load)}$$

[a] Production output is an example of an independent variable that commonly impacts manufacturing process energy use.
[b] Use the most recent typical meteorological year dataset. As of January 2014, the most comprehensive national typical meteorological year dataset is TMY3. Evaluators should confer with the local jurisdiction to see if they should use a different, regional, dataset instead.

4.4.2 Auxiliary Equipment

Table 4 lists additional steps for determining the auxiliary savings term in Equation 1 (kWh SavingsTotal = kWh SavingsChiller + kWh SavingsAuxiliary).

Table 4. Auxiliary Equipment M&V Procedures

Step	Details
Measure baseline[a] and reporting period auxiliary demand data	If the energy consumption of auxiliary equipment is constant, take spot measurements on the auxiliary equipment affected by the chiller measure. If consumption of auxiliary equipment is variable *and* the chiller plant is serving an HVAC load, take short-term measurements at representative load levels for auxiliary equipment affected by the chiller measure. If consumption of auxiliary equipment is variable *and* the chiller plant is serving a process load, take continuous measurements over the length of each type of process cycle for all auxiliary equipment affected by the chiller measure. If more than one piece of auxiliary equipment is affected, the measurements across affected equipment should be coincident.
Develop bin data and sum the kilowatt-hour savings	Bin baseline and reporting period data using bin profiles established for the chiller (if consumption of auxiliary equipment is constant—as it might likely be for the baseline scenario; kilowatts will be the same for all bins). Calculate kilowatt-hour savings by bin and sum as described in Table 3.

[a] If auxiliary equipment is replaced as part of a replace-on-burnout or natural turnover project, the building code could require upgrades to the auxiliary equipment. If this is the case, establish a hypothetical baseline for the affected auxiliary equipment.

4.5 Regression Modeling Direction

Calculating normalized savings for the majority of projects—whether following the IPMVP's Option A or Option C—will require the development of a baseline and reporting period regression model.[16] Use one of the following three types of analysis methods to create the model:

- **Linear regression:** For one routinely varying significant parameter (e.g., OAT).[17]
- **Multivariable linear regression:** For more than one routinely varying significant parameter (e.g., OAT, process parameter).
- **Advanced regression:** For a multivariable, nonlinear fit requiring a polynomial or exponential model.[18]

[16] This could either be a single regression model that uses a dummy variable to differentiate the baseline/reporting period data or two independent models for the baseline and reporting period, respectively.
[17] One of the most common linear regression models is the three-parameter change point model. For example, a model that represents cooling electricity consumption will have one regression coefficient that describes non-weather-dependent electricity use, a second regression coefficient that describes the rate of increase of electricity use with increasing temperature, and a third parameter that describes the change point temperature, also known as the balance point temperature, where weather-dependent electricity use begins.
[18] Evaluators may need to use advanced regression methods if a chiller plant is providing cooling for manufacturing or industrial processes.

Develop all models in accordance with common practices and only use them when statistically valid (see Section 4.5.2, *Testing Model Validity*). If there are no significant independent variables (as would be the case for a constant-process cooling load), evaluators are not required to use a model because the calculated savings are inherently normalized.

4.5.1 Recommended Method for Model Development

Use cooling-load data and independent-variable data that are representative of a full cycle of operation to the maximum extent possible. For example, if a chiller plant located in New England is serving an HVAC load with a temperature adjustment during unoccupied hours, then collect load data across the full range of outdoor air temperatures for each of the operating schedules (occupied and unoccupied) for each season. Table 5 provides an example of the data required for model development.

Table 5. Example of Data Required for Model Development

	Shoulder Season	Summer Season
Occupied Hours	Short-term load measurements during occupied hours. Measurements should be representative of full range of shoulder season OAT (approximately 10 OAT bins).	Short-term load measurements during occupied hours. Measurements should be representative of full range of summer season OAT (approximately 10 OAT bins).
Unoccupied Hours	Short-term load measurements during unoccupied hours. Measurements should be representative of full range of shoulder season OAT (approximately 10 OAT bins).	Short-term load measurements during unoccupied hours. Measurements should be representative of full range of summer season OAT (approximately 10 OAT bins).

Analyze the data collected to identify outliers. Only remove outliers when there is a tangible explanation to support the erratic data points. Discussion of how to identify outliers is outside the scope of this protocol.

4.5.2 Testing Model Validity

To assess the accuracy of the model, begin by reviewing the parameters listed in Table 6 (EVO 2012).

Table 6. Model Statistical Validity Guide

Parameter Evaluated	Description	Suggested Acceptable Values
Coefficient of determination (R^2)	A measure of the extent to which the regression model explains variations in the dependent variable from its mean value.	> 0.75
T-statistic (absolute value)	An indication of whether the regression model coefficients are statistically significant.	> 2[a]
Mean bias error	An indication of whether the regression model overstates or understates the actual cooling load.	Will depend on the project, but generally: <± 5%

[a] Determine the t-statistic threshold based on the evaluator's chosen confidence level; a 95% confidence level requires a t-statistic of 1.96. Evaluators should determine an acceptable confidence level depending on project risk (i.e., savings risk), budget, and other considerations.

A model outside the suggested range indicates parameter coefficients that are relatively poorly determined, with the result that normalized consumption will have relatively high statistical prediction error. Ordinarily, evaluators should not use such a model for normalization, unless the analysis includes appropriate statistical treatment of this prediction error. Discussion of how to proceed in such circumstances is outside the scope of this protocol.

When possible, attempt to enhance the regression model by:

- Increasing or shifting the measurement period
- Incorporating more data points
- Including independent variables previously unidentified
- Eliminating statistically insignificant independent variables.

Also, when assessing model validity, consider the coefficient of variation (CV) of the root mean squared error (RMSE), fractional savings uncertainty, and residual plots. Refer to ASHRAE Guideline 14-2002 and Bonneville Power Administration's *Regression for M&V: Reference Guide* for direction on how assess these additional parameters.

5 Sample Design

Consult the Uniform Methods Project's *Chapter 11: Sample Design Cross-Cutting Protocol* for general sampling procedures if the chiller project population is sufficiently large or if the evaluation budget is constrained. Ideally, use stratified sampling to partition chiller projects by facility type, process vs. HVAC load, and/or the magnitude of claimed (*ex ante*) project savings. Stratification ensures evaluators can confidently extrapolate sample findings to the remaining project population. Regulatory or program administrator specifications typically govern the confidence and precision targets, which will influence sample size.

6 Other Evaluation Issues

When claiming lifetime and net program chiller measure impacts, consider the following evaluation issues in addition to first-year gross impact findings:

- Net-to-gross estimation
- Early replacement
- Dual baseline realization rates.

6.1 Net-to-Gross Estimation

The Uniform Methods Project's cross-cutting *Estimating Net Savings: Common Practices* discusses an approach for determining net program impacts at a general level. It is recommended that the collection between gross and net impact results and teams collecting site-specific impact data to ensure there is no double counting of adjustments to impacts at a population level.

6.2 Early Replacement

As a supplement to the Uniform Methods Project's *Estimating Net Savings: Common Practices*, the evaluator should consider assessing whether early replacement projects were program-induced. If the early replacement was not program-induced, it is appropriate to use a hypothetical baseline rather than a dual baseline.

6.3 Dual-Baseline Realization Rates

For program-induced early replacement projects, two different realization rates (evaluated [*ex post*] gross savings/claimed [*ex ante*] gross savings) exist over the EUL of the new equipment:

- **Period 1 Realization Rate.** The realization rate is applicable over the first part of the dual baseline; evaluators should calculate the gross *ex post* savings using the existing equipment as the baseline.
- **Period 2 Realization Rate.** The realization rate is applicable over second part of the dual baseline; evaluators should calculate the gross *ex post* savings using a hypothetical baseline.

Therefore, if reporting life cycle gross impact findings, evaluators need to account for both Period 1 and Period 2 realization rates.

7 References

ASHRAE. (2002). Guideline 14-2002: Measurement of Energy and Demand Savings. American Society of Heating, Refrigeration and Air-Conditioning Engineers.

ASHRAE. (2012). *2012 ASHRAE Handbook: HVAC Systems and Equipment.* American Society of Heating, Refrigeration and Air-Conditioning Engineers.

CPUC. (2008). *Database for Energy Efficiency Resources (DEER)*. California Public Utilities Commission.

EVO. (January 2012). *International Performance Measurement and Verification Protocol – Concepts and Options for Determining Energy and Water Savings Volume 1.* Efficiency Valuation Organization.

Fagan, F., Bradley, K., Lutz, A. (2011). *Strategies for Improving the Accuracy of Industrial Program Savings Estimates and Increasing Overall Program Influence*. Prepared for the International Energy Program Evaluation Conference.

Ridge, R., Jacobs, P., Tress, H., Hall, N., Evans, B. (2011). *One Solution to Capturing the Benefits of Early Replacement: When Approximately Correct is Good Enough*. Prepared for the International Energy Program Evaluation Conference.

US DOE FEMP. (April 2008). *M&V Guidelines: Measurement and Verification for Federal Energy Projects—Version 3.0*. U.S. Department of Energy Federal Energy Management Program.

8 Bibliography

BPA. (2011). *Regression for M&V: Reference Guide.* Bonneville Power Administration.

CPUC. (April 2006). *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*. California Public Utilities Commission.

Connecticut Light & Power, United Illuminating Company. (2005). *UI and CL&P Program Savings Documentation for 2006 Program Year*.

Itron, Inc. (February 2010). *2006-2008 Evaluation Report for PG&E Fabrication, Process and Manufacturing Contract Group*. Prepared for California Public Utilities Commission.

Jacobson, D., Studer, E. (2007). *Lessons Learned from a Decade of Evaluating Customized Commercial and Industrial Efficiency Measures*. Prepared for the International Energy Program Evaluation Conference.

NEEP. (May 2010). *Regional EM&V Methods and Savings Assumptions Guidelines.* Northeast Energy Efficiency Partnerships.

OPA. (March 2011). *saveONenergy Retrofit Program Project Measurement and Verification Procedures.* Ontario Power Authority.

RLW Analytics, Inc. (October 2008). *2004-2005 Los Angeles County —Internal Services Department / Southern California Edison/Southern California Gas Company Energy Efficiency Partnership Impact Evaluation Study*. Prepared for California Public Utilities Commission.

Robert Mowris & Associates. (June 2005). *Measurement & Verification Load Impact Study for NCPA SB5X Commercial and Industrial HVAC Incentive Programs*. Prepared for Northern California Power Agency.

Summit Blue Consulting, LLC. (September 2008). *APS Measurement, Evaluation & Research (MER) Report—Solutions for Business*. Prepared for Arizona Power Service.

Chapter 15: Commercial New Construction Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in September 2014.

Steven Keates
ADM Associates, Inc.
Sacramento, California

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68571
September 2017

Contract No. DE-AC36-08GO28308

Chapter 15: Commercial New Construction Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in September 2014.

Steven Keates
ADM Associates, Inc.
Sacramento, California

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68571
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

These methods, processes, or best practices (“Practices”) are provided by the National Renewable Energy Laboratory (“NREL”), which is operated by the Alliance for Sustainable Energy LLC (“Alliance”) for the U.S. Department of Energy (the “DOE”).

It is recognized that disclosure of these Practices is provided under the following conditions and warnings: (1) these Practices have been prepared for reference purposes only; (2) these Practices consist of or are based on estimates or assumptions made on a best-efforts basis, based upon present expectations; and (3) these Practices were prepared with existing information and are subject to change without notice.

The user understands that DOE/NREL/ALLIANCE are not obligated to provide the user with any support, consulting, training or assistance of any kind with regard to the use of the Practices or to provide the user with any updates, revisions or new versions thereof. DOE, NREL, and ALLIANCE do not guarantee or endorse any results generated by use of the Practices, and user is entirely responsible for the results and any reliance on the results or the Practices in general.

USER AGREES TO INDEMNIFY DOE/NREL/ALLIANCE AND ITS SUBSIDIARIES, AFFILIATES, OFFICERS, AGENTS, AND EMPLOYEES AGAINST ANY CLAIM OR DEMAND, INCLUDING REASONABLE ATTORNEYS' FEES, RELATED TO USER’S USE OF THE PRACTICES. THE PRACTICES ARE PROVIDED BY DOE/NREL/ALLIANCE "AS IS," AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL DOE/NREL/ALLIANCE BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER, INCLUDING BUT NOT LIMITED TO CLAIMS ASSOCIATED WITH THE LOSS OF PROFITS, THAT MAY RESULT FROM AN ACTION IN CONTRACT, NEGLIGENCE OR OTHER TORTIOUS CLAIM THAT ARISES OUT OF OR IN CONNECTION WITH THE ACCESS, USE OR PERFORMANCE OF THE PRACTICES.

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter author wishes to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Taghi Alereza of ADM Associates, Inc.
- Joshua Taylor of ADM Associates, Inc.

Suggested Citation

Keates, Steven. (2017). *Chapter 15: Commercial New Construction Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68571. www.nrel.gov/docs/fy17osti/68571.pdf

Acronyms

ANSI	American National Standards Institute
CV	coefficient of variation
CVRMSE	coefficient of variation of the root mean square error
DSM	demand-side management
ECM	energy conservation measure
EM&V	Evaluation, measurement, and verification
HVAC	heating, ventilating, and air conditioning
IPMVP	International Performance Measurement and Verification Protocol
LEED	Leadership in Energy & Environmental Design
M&V	monitoring and verification
NMBE	normalized mean bias error
TMY	typical meteorological year

Table of Contents

List of Figures

List of Tables

1 Measure Description

This protocol is intended to describe the recommended method when evaluating the whole-building performance of new construction projects in the commercial sector. The protocol focuses on energy conservation measures (ECMs) or packages of measures where evaluators can analyze impacts using building simulation. These ECMs typically require the use of calibrated building simulations under Option D of the International Performance Measurement and Verification Protocol (IPMVP).[1]

Examples of such measures include Leadership in Energy & Environmental Design (LEED) building certification, novel and/or efficient heating, ventilation, and air conditioning (HVAC) system designs, and extensive building controls systems. In general, it is best to evaluate any ECM expected to significantly interact with other systems within the building and with savings sensitive to seasonal variations in weather.[2] The protocol classifies commercial new construction projects as:

- **Newly constructed buildings:** The design and construction of an entirely new structure on a greenfield site or wholesale replacement of a structure torn down to the ground.
- **Addition (expansion) to existing buildings:** Significant extensions to an existing structure that requires building permits and triggers compliance with current codes.
- **Major renovations or tenant improvements of existing buildings:** Significant reconstruction or “gut rehab” of an existing structure that requires building permits and triggers compliance with current codes.

Evaluators may need to apply the evaluation methods described here for new construction projects for some projects in the retrofit programs. While some retrofit projects have much in common with new construction projects, their scope does not uniformly fall under the new construction categories previously described. Evaluators should assess these projects according to the guidelines described for retrofit equipment (described in separate protocols).

Evaluation, measurement, and verification (EM&V) of new construction programs involves unique challenges, particularly when defining baseline energy performance. An agreed-upon building energy code or industry standard defines the baseline equipment evaluators use to measure energy impacts for new construction measures. As the baseline equipment for new construction measures does not physically exist and cannot be measured or monitored, evaluators typically employ a simulation approach. Due to the nuances involved in appropriately determining baseline equipment/performance evaluations, experienced professionals with a good understanding of building construction practices, simulation code limitations, and the relevant building codes should oversee these types of projects.

---

[1] As discussed in the section “Considering Resource Constraints” of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

[2] Note the term whole-building modeling does not necessitate use of sophisticated stand-alone simulation software (e.g., eQUEST, EnergyPlus). It is acceptable to employ engineering models using spreadsheet calculations, provided they meet the guidelines set forth in Section 4.

Further, evaluators typically assess new construction measures within the first few years of construction. During this period, there is often considerable change in building occupancy and operation before the measures design intent becomes realized. This results in additional challenges for evaluators using monitored data and/or facility utility billing or energy consumption history to define as-built building performance.

2 Application Conditions of Protocol

Use the algorithms and protocols described here to evaluate new construction whole-building performance ECMs installed in commercial facilities. When new construction ECMs do not directly impact HVAC energy use, it is often possible to use spot measurements and engineering calculations to evaluate savings with sufficient rigor (ASHRAE 2002). This is usually the case, for example, with lighting and domestic hot water retrofits.[3] This protocol does not cover the guidelines for selecting the appropriate monitoring and verification (M&V) rigor for such measures. Consult the IPMVP or measure-specific protocols within the Uniform Methods Project protocols to review evaluation guidelines for measures that do not require calibrated building simulation.

2.1 Incentive Types

Program administrators typically classify new construction demand-side management (DSM) program incentives as being either component-based or performance-based and design the program to offer one or both types of incentives.

2.1.1 Component-Based Incentives

Component-based (or "prescriptive") incentives tend to involve individual technologies and equipment. Examples of prescriptive incentives may include lighting fixtures, occupancy sensors, motors, and small packaged (unitary) HVAC units. Evaluators often determine rebate amounts and claimed savings estimates based on stipulated per-unit estimates.[4] Evaluators will sometimes assess component-based rebates according to measure-specific protocols using partial or complete retrofit isolation evaluation strategies (IPMVP Option A or Option B).

2.1.2 Performance-Based Incentives

Performance-based incentives tend to target more complex projects involving improvements to the overall building energy performance.

Whole-building performance incentives can:

- Encompass various specific (above-code) upgrades
- Fund design, analysis, equipment, and/or installation (labor) costs.[5]

An example of a performance-based project is LEED certification. Buildings that are LEED certified often encompass ECMs that range from envelope improvements to high-efficiency equipment installations (often going beyond just HVAC) and complicated controls algorithms.

---

[3] While the general magnitude of the secondary impacts imparted by lighting measures on HVAC equipment are well-established for various building types, take care to estimate these impacts appropriately in new construction building stock. New buildings typically have more efficient HVAC equipment, which reduces the magnitude of heating and cooling interactive effects. Secondary impacts can be estimated using prototypical building models, representative of the physical facility. See the Uniform Method Project's *Chapter 2: Commercial and Industrial Lighting Evaluation Protocol* or CPUC 2004 for guidelines regarding HVAC interactive factors.

[4] Units used do not necessarily represent quantity. Frequently applied units include: installed horsepower, tons of refrigeration, and square footage.

[5] Some new construction programs have been successfully implemented without direct financial incentives (design assistance, financing, etc.).

The complex interactions between these ECMs can only be reliably determined through the use of calibrated building simulation models.

Performance-based incentive amounts are typically determined by the expected annual energy and/or demand impacts (e.g., per kilowatt-hour, therm, kilowatt).[6] Annual energy-savings estimates for performance-based projects (and programs) require evaluators to use custom calculations via whole-building simulation modeling tools. Therefore, highly skilled technical labor is required to successfully implement and evaluate these programs.[7]

[6] Depending on program design, the "expected" energy impacts can be either *ex ante* or *ex post*.
[7] See Johnson & Nadel 2000 for more information.

3 Savings Calculations

Use the following algorithm to calculate energy savings for new construction measures. Note that evaluators can calculate demand savings using the same algorithms by simply substituting "demand" for "energy use."[8]

Equation 1

Energy Savings = Projected Baseline Energy Use – Post-construction Energy Use

Where,

Projected Baseline Energy Use	= Projected energy use of baseline systems at full design occupancy and typical building operating conditions
Post-construction Energy Use	= Energy use of measure systems at full design occupancy and typical building operating conditions

As described in Section 4, *Measurement and Verification Plan*, calculate projected baseline energy use and post-construction energy use using a whole-building simulation model that is calibrated to monthly (or hourly) utility energy consumption histories. Evaluators can use four components to report savings for new construction ECMs:

- Expected (planned) measure savings
- Rebated measure savings
- Non-rebated measure savings
- Total achieved savings

Section 4 discusses each component.

[8] When calculating the coincident peak demand savings, average the hourly demand savings over the "peak demand window" period, as defined by the utility.

4 Measurement and Verification Plan

4.1 International Performance Measurement and Verification Protocol Option

The preferred approach to calculate savings for whole-building performance new construction projects is calibrated building simulation models according to IPMVP Option D (IPMVP 2006). The recommended approach requires sufficient resources be allocated to the project to allow for detailed onsite data collection, preparation of the simulation models, and careful calibration. The method is less costly when a functioning *ex-ante* model is available to the evaluator, though obtaining the *ex-ante* model is not a prerequisite to its application.

Determine the appropriate modeling software by the specifics of the evaluated buildings (e.g., HVAC system and zoning complexity, building constructions, complexity of the ECMs); there is no single software (currently available) that can simulate all variations of HVAC system types, building constructions, and ECMs. Thus, it may be necessary to use multiple tools to evaluate building performance accurately.

In general, the appropriate software for modeling building systems and energy performance must:

- Create outputs that comply with American National Standards Institute (ANSI)/ASHRAE Standard 140-2011[9]
- Accurately simulate the building's systems and controls
- Use an hourly or sub-hourly time step to perform simulation[10]
- Simulate building performance using user-defined weather data at hourly intervals

For more information on specific requirements for simulation software, see pp. 133 in *The California Evaluation Framework* (CPUC 2004) and pp. 26-27 in *Appendix J – Quality Assurance for Statistical, Engineering, and Self-Report for Estimating DSM Program Impacts* (CADMAC 1998).[11]

The U.S. Department of Energy's (DOE) Energy Efficiency and Renewable Energy website[12] contains a list of building energy simulation software. Although some tools listed are proprietary, the website also lists public-domain DOE-sponsored tools. Summary comparisons and descriptions of commonly used software can be found in Crawley (2005).

The preferred full Option D approach will in some cases be intractable due to limited data availability or evaluation budgetary limitations. In such cases, alternate methodologies are acceptable but the following guidelines should be followed:

---

[9] ANSI/ASHRAE Standard 140-2011 establishes test procedures validating software used to evaluate thermal performance of buildings (and applicable HVAC equipment).

[10] It is preferable the software use unique time steps for each interval (e.g., 8,760 hours).

[11] For further commentary on simulation software requirements, see ASHRAE 2002, IPMVP 2001, and IPMVP 2006.

[12] The DOE's Energy Efficiency and Renewable Energy website can be found at: http://apps1.eere.energy.gov/buildings/tools_directory/.

- Onsite verification and review of as-built drawings and commissioning reports (as available) should be performed to verify which energy saving features were actually installed and are functioning

- Ex-ante savings calculations should be based in a whole building simulation model of the building or of a building that is representative of the actual facility

- Results should be compared with billing data (when available), engineering rules of thumb, and/or secondary literature to review reasonability.

4.1.1 Verification Process

Figure 1 depicts the overall process to verify savings under Option D, from *The California Evaluation Framework* (CPUC 2004). The process starts by specifying which site data collection and equipment monitoring requirements are in an M&V plan. Additionally, the M&V plan should specify:

- The applicable version of the building codes and equipment standards that determine the baseline (or applicable 'practice' that may determine baseline). This is discussed in more detail in Section 4.3.
- The above-code technologies present in the building (claimed as ECMs)
- The software for modeling building performance
- Appropriate data for calibrating the simulations
- How to address modeling uncertainties
- Against what statistical indices calibration will be measured.

While reviewing the energy consumption data can be useful in developing data collection needs, it is not a prerequisite to creating and implementing the M&V plan. However, when developing the M&V plan, evaluators should consider how long a building has been occupied because that will determine amount and granularity of energy consumption data available. Fewer months of consumption data, or the availability of only monthly data, usually means there will be a greater emphasis on metering specific pieces of equipment. Conversely, the presence of a building automation system, energy monitoring system, lighting control panels, (collectively referred to here as building automation system) or other devices to control and/or store data about the operational characteristics of the building will allow for a lesser dependence upon utility usage data.

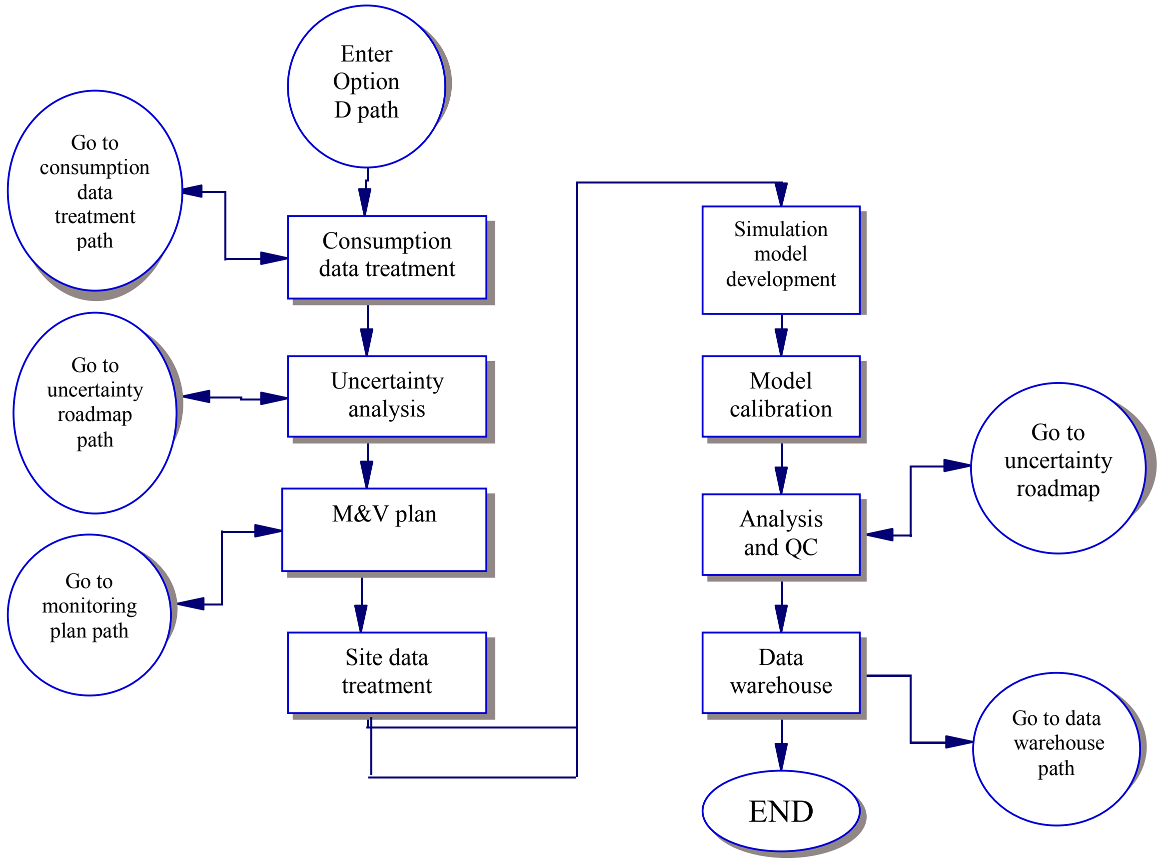


Figure 1. Roadmap for IPMVP Option D

4.1.2 Data Requirements and Collection Methods

Data collected during this step includes all of the information required to define and calibrate the building simulation model. Due to the unique nature of each new construction project, it is impractical to prescribe a comprehensive list of specific parameters evaluators should collect on site. Instead, use the following guidelines to identify key data points and minimize the uncertainty in the final calibrated simulations. After identifying specific parameters, refer to the Uniform Methods Project's *Metering Cross-Cutting Protocols* for instructions regarding the methods to submeter the physical parameters.

The data used to define building simulation models come from stipulated and physical sources. Furthermore, these data can be static or dynamic in nature, as described here:

- *Static data points.* These are essentially constant values that describe physical properties of the equipment and the building surfaces or the set point and operational range controlling the building equipment. [13] Examples of static data points are window glazing, motor efficiencies, and thermostat set points.

[13] Set points can refer to a control zone, thermostat, control valve, flow rate, voltage, photocell, or other parameter that is designed to maintain optimal environmental conditions within the building. Some set points are "dynamic" in that they may change according to the time of day.

- *Dynamic data.* These are time-dependent variables that describe building and equipment operations. These data capture the behavioral and operational details (e.g., weather, motor loading, and building occupancy) needed to establish a building's energy-use characteristics. Dynamic data, which are often the most difficult to collect, represent the greatest source of uncertainty in a building simulation.

IPMVP Option D (IPMVP 2006) allows use of stipulated data, although it is important to minimize the number of these inputs, as they represent degrees of freedom (and, therefore, additional uncertainty) in the model. Sources for such data include peer-reviewed research, engineering references, simulation program defaults, manufacturers' specifications, and/or survey information from on-site visits (e.g., mechanical and architectural drawings and visual inspection of nameplate information).

The following are convenient categories of important physical data to collect on site (ASHRAE 2002):

- Lighting systems
- Plug loads
- HVAC systems
- Building envelope and thermal mass
- Building occupants
- Other major energy-using loads.[14]

Another important element of the data collection process entails the use of submetering to define behavioral and dynamic aspects of a building and its subsystems. In this protocol, the term submetering encompasses both direct placement of monitoring equipment by evaluation personnel and collecting data from the building automation systems (also known as trend data) when available. Even when the absolute accuracy of the collected data is unknown, submetered data is useful for informing operational schedules (e.g., lighting and ventilation) and calibrating the model.

The degree of submetering required is largely dependent upon the quality and resolution of the facility's energy consumption history. The following descriptions of submetering represent the minimum amount of data collected for calibrating simulation models. Additional submetering may be necessary to verify complex control schemes and/or set points. Perform additional submetering as budget and time permit.[15] Use such data to inform model inputs rather than to function as a calibration target.

[14] This category is particularly important in buildings such as grocery stores, refrigerated warehouses, and some retail.
[15] For example, verifying functionality of chilled water reset controls or condensing water relief set points.

4.1.2.1 Submetering With Monthly Bills

When only a monthly utility billing history is available for a facility, it is important to submeter both HVAC fan schedules[16] and interior lighting fixtures. Also, if the facility has unique or considerable equipment loads (e.g., data centers), meter these as well.

When monitoring unitary HVAC equipment, isolate the power used by fans from that used by compressors. This ensures evaluators can use the resulting data when calibrating time-of-use and magnitude of fan power.

If, due to site or budget limitations, the electrical monitoring must comprise the unitary system as a whole, use motor nameplate information and fan curves in conjunction with local weather data to disaggregate the fan and compressor power.[17]

Alternatively, use one-time power measurements to establish a unit's demand for each operation mode. Combine these measurements with time-series data to identify time spent in each operation mode and, thereby, determine the fan schedules.

4.1.2.2 Submetering With Hourly Bills

Hourly (or sub-hourly) energy consumption histories contain much more information for model calibration than monthly usage alone. While this additional information reduces submetering requirements, it does not eliminate the need to submeter HVAC fan schedules as they are important for disaggregating base loads from ventilation. As described for monthly billing data, consider submetering other large energy-using features (e.g., pool-heating and space-cooling equipment, atria lighting, and internet technology loads) if possible given evaluation budgets.

4.2 Simulation Model Development

It is important to model several iterations of the simulated building so as to fully capture the various aspects of the savings for new construction ECMs. Table 1 lists this iterative process, which entails three versions of the as-built building and two versions of the baseline building, including:

- As-built physical
- As-built design
- As-built expected design
- Whole-building reference
- Measure building reference.

Table 1 does not include intermediate modeling of individual ECMs. Intermediate modeling can be used to disaggregate individual measure impacts and interactive effects. If measure-level

[16] It is important to capture a building's ventilation schedule when HVAC systems are used to supply outside air to maintain required fresh requirements. If performing submetering on a sample of HVAC fans, place priority on accurately capturing when (and how much) outside air is introduced into the building.

[17] To employ this method, the modeler must have the requisite expertise to apply appropriate statistical and engineering modeling techniques to perform this analysis. For further information on energy consumption analysis, see the *Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol*.

savings estimates (and therefore, intermediate modeling of measures) is required, work with the governing jurisdiction for the evaluation process to establish an appropriate hierarchy to govern the order in which measures are stacked and individual measure savings assessed.

Table 1. List of Models Used To Simulate Savings for New Construction ECMs

Model	Model Name and Purpose	Model Description
1	As-Built Physical *To calibrate simulations and assess uncertainty*	Model and simulate, as found during site visit. Use the occupancy and building operation, as reflected in billed energy history and submetered data. Simulate using actual local weather observations matching the consumption history period.
2	As-Built Design *To estimate typical usage at full occupancy*	Base on as-built physical model. Use full design occupancy and expected typical building schedules. Use construction and equipment efficiencies, as found during site visits. Simulate using normalized weather data (e.g., typical meteorological year [TMY] datasets).[a]
3	As-Built Expected Design *To estimate difference between original and as-built models*	Base on as-built design model. Use full design occupancy and expected typical building schedules. Use assumed constructions and equipment efficiencies. Simulate using normalized weather data (e.g., TMY datasets).
4	Whole-Building Reference *To estimate savings of the ECMs*	Base on as-built design model. Use full design occupancy and expected typical building schedules. Apply baseline requirements defined by reference codes or standards. Simulate using normalized weather data (e.g., TMY).
5	Measure Building Reference *To isolate savings claimed by the participant*	Base on whole-building reference model. Use full design occupancy and expected typical building schedules. Apply baseline requirements defined by reference codes or standards. Include ECMs not incentivized by DSM program. Simulate using normalized weather data (e.g., TMY).

[a] Note the TMY are referenced here as an example series of normalized weather data. When incorporating TMY weather data, use TMY3 weather data when available. While TMY weather represents a common standard, review the reporting needs of the project, as other normalized weather datasets may be more appropriate (e.g. Weather year for Energy Calculations [WYEC] or California Thermal Zones [CTZ]).

Begin the development of the model by generating a model of the building as it was built and is operating during the site visit—and as reflected by utility energy consumption data. Use this initial model, the as-built physical model, to calibrate the modeled building to available physical data. This ensures evaluators can use successive iterations in a predictive capacity. A detailed discussion of the calibration process falls outside the scope of this protocol; however, for detailed calibration procedures and guidelines see Section 6.3.3.4 in ASHRAE Guideline 14-2002 (ASHRAE 2002).

Once calibrated, use the as-built physical model to generate the as-built design model, which should reflect the building at full-design occupancy and operation according to expected typical schedules. The only differences between these models are building occupancy, operational schedules, and any modeling guidelines incorporated from codes or standards used to define baseline performance. For buildings currently operating at full occupancy, there may be very little difference between these models. Refer to Tables 11.3.1 and G3.1 in ASHRAE Standard 90.1-2007 (ASHRAE 2007) for examples of modeling requirements specified by codes and standards.

Then, use the as-built design model to generate the as-built expected design model. While this model simulates the building's operation according to its design intent, it also includes claimed assumptions regarding envelope constructions and equipment efficiencies. Review the model for discrepancies between claimed assumptions and the physical building; if no discrepancies exist, this model will be identical to the as-built design.

After developing as-built models, evaluators can model baseline building performance, which results in the whole-building reference model; to generate this model, apply the appropriate codes and standards used to define baseline building performance to the as-built design model. The M&V plan should identify such standards before modeling begins. The following section, *Baseline Considerations*, discusses additional considerations for baseline selection. Similar to the as-built design model, the whole-building reference model should reflect the building's operation according to its expected long-term patterns while using equipment and construction that minimally complies with the reference code or standard.

Finally, start with the whole-building reference model to generate the measure building reference model—this model will include ECMs not incentivized by the DSM program. It is likely all the implemented ECMs are included in the whole-building performance incentives; therefore, both the baseline models may be identical. However, as incentives often are applied for during the building's design and construction process, additional above-code equipment or construction may be implemented that were not included in the final incentive.

4.3 Baseline Considerations

Defining baseline building physical characteristics and equipment performance is one of the most important (and difficult) tasks in evaluating savings for new construction ECMs. This is for several reasons. As noted, new construction ECMs do not have a physical baseline to observe, measure, or document. Rather, evaluators must define the baseline "hypothetically" through an appropriate interpretation of the applicable energy codes and standards. It is typically complicated to establish an *appropriate* interpretation due to the overlapping scope of federal, state, and local codes. Conversely, some states do not have a building energy-efficiency standard separate from the federal standards. Typically, evaluators determine baseline building characteristics and equipment performance requirements by locally adopted building energy codes. In some cases, however, applying a more rigorous, above-code baseline may better reflect standard local construction or industry-standard practices. Thus, in addition to a good understanding of the relationship between federal, state, and local standards, evaluators may need to consult with program guidelines (which often specify greater than code stringency or other technical specifications) or statewide evaluation frameworks.

Enforcement of the state codes is the responsibility of the local building officials. The EM&V effort of energy-efficiency programs is usually carried out by utility or other program administrators or by a public utilities commission. Whereas the public utilities commission usually has no enforcement responsibility for the codes and standards, they often point to the official state standards as the governing document regardless of the degree of enforcement of those codes at the local level.

In general, the baseline must satisfy the following criteria (IPMVP 2006):

- It must appropriately reflect how a contemporary, nonparticipant building would be built in the program's absence.[18]
- Evaluators must rigorously define it with sufficient detail to prescribe baseline conditions for each individual ECM and for the building components simulated.
- Evaluators must develop it with sufficient clarity and documentation to be repeatable.

The BCAP-OCEAN website (http://energycodesocean.org) can be a useful resource in identifying locally adopted energy codes and standards when starting the evaluation of a whole-building or commercial new construction project.

4.4 Calculating Savings

To calculate savings, apply simulation outputs (from models 2 through 5 in Table 2) to the formulas described in Section 3. In all cases except as-built physical, simulate the post-construction energy use and the projected baseline energy use using normalized weather data (TMY).

As discussed in Section 3, there are four components that comprise calculated energy savings (defined in Table 2 and shown in Figure 2). Determine the final reported (verified) savings values in the context of M&V objectives.

[18] Locally adopted building codes will define gross savings of new construction programs. Only consider standard construction practices of nonparticipant buildings when performing a net-to-gross analysis. One notable exception is when the evaluated program defines its own baseline, according to an above-code standard (for example, ASHRAE Standard 189.1-2011).

Table 2. Comparison of Savings Components for New Construction ECMs

Savings Component	Model Subtraction	Description
Expected Measure Savings	N/A	Energy savings expected by the building designers and/or the DSM program application (also known as the project's planned energy savings).
Rebated Measure Savings	5 – 2	Evaluated (or realized) energy savings for incentivized ECMs, often determined by an independent third-party evaluator. Calculate these savings by subtracting the difference in simulated energy use of the as-built design from the measure building reference (the result is also known as the project's *ex post* savings).
Nonrebated Measure Savings	4 – 5	Energy savings resulting from ECMs implemented in the final building design, but not rebated by the DSM program. Calculate these savings by subtracting the difference in simulated energy use of the measure building reference from the whole-building reference (the result is also known as the spillover savings).
Total Achieved Savings	4 – 2	Evaluated (or realized) energy savings for all implemented ECMs, whether rebated or not. These are often determined using an independent third-party evaluator, and calculated by subtracting the difference in simulated energy use of the as-built design from the whole-building reference. Some DSM programs report this (rather than rebated measure savings) as the project's *ex post* savings.

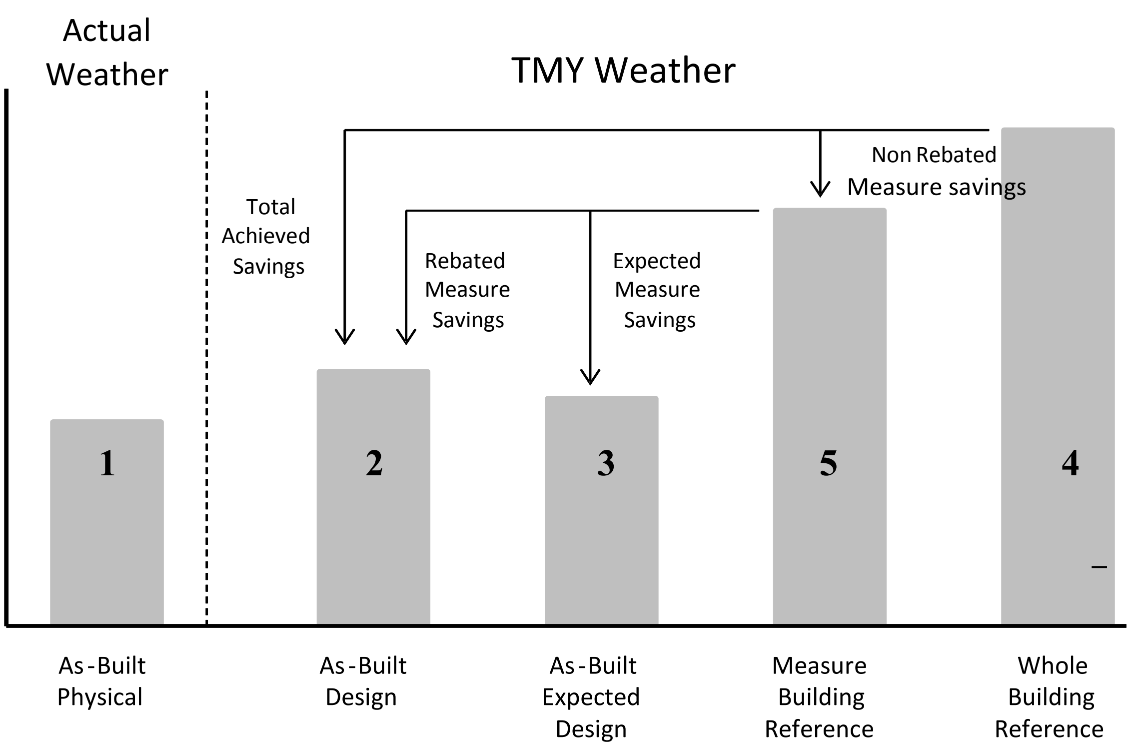


Figure 2. Illustration of savings components for new construction ECMs

4.5 Quantify and Locate Modeling Uncertainty

Due to the complex set of physical, thermodynamic, and behavioral processes simulated, it is difficult to fully characterize the uncertainty in modeled outputs without multiple statistical and analytical tools. Additionally, practical limitations on budgets and time allotted for M&V activities frequently result in qualifying uncertainty in final simulated savings by reporting uncertainty in the model's calibration to energy consumption history. Quantify calibration uncertainty using the normalized mean bias error (NMBE) and coefficient of variation of the root mean square error (CVRMSE).[19] Pages 13-16 of ASHRAE Guideline 14-2002 (ASHAE 2002), provides detailed descriptions of these calculations and their applications.

Determine calibration uncertainty by comparing outputs from the calibrated as-built physical model with the facility's consumption history. Table 3 shows calibration uncertainty targets for monthly and hourly consumption history resolutions (ASHRAE 2002).

Table 3. Acceptable Tolerances for Uncertainty in Calibrated Building Simulations

Resolution of Energy Consumption History	NMBE Tolerance	CVRSME Tolerance
Monthly	±5%	±15%
Hourly	±10%	±30%

As newly constructed buildings have a short energy consumption history, it is important to consider how many monthly observations are required to attain a suitably calibrated model. The amount of consumption history required for calibration depends on building type and occupancy. Buildings with little seasonal variations in energy use[20] and short ramp-up periods may need as little as three or four months of consumption history, assuming building occupancy and usage are well-defined and stable. Typically, buildings in this category include grocery stores, restaurants, and data centers.

Conversely, buildings that experience significant seasonal variation, or that are not fully occupied for extended periods, may require a complete year (or more) of consumption history before modelers can determine a reliable calibration. For these buildings, occupancy and usage must be well-defined and stable during all observations used for calibration. Typical buildings of this type include offices, schools, and malls (both strip and enclosed).

Mandating definitive requirements for the minimum number of observations required to sufficiently calibrate a simulation would unduly constrain modelers and could place impractical limitations on EM&V efforts. However, this protocol recommends the following as guidelines:

- Observations should sufficiently characterize a building's energy use, so modelers can extrapolate reliable annual energy-use values.

[19] These two statistical measurements provide an assessment of the variance between the simulated and measured (by the utility meter) energy use and electric demand. This protocol considers modeling uncertainty acceptable when this variance is below the thresholds suggested in Table 3.

[20] Although energy used by HVAC systems can vary seasonally, such usage generally correlates well with outside weather. Thus, the energy simulation model can sufficiently extrapolate such seasonality (when simulated using the appropriate weather data), reducing the number of billed observations required to calibrate buildings having HVAC use that is dominated by weather.

- Observations should sufficiently describe expected seasonal variations in building operations.
- Building occupancy and operating conditions must be known for the set of observations.
- Building occupancy and operating conditions must remain stable for the duration of observations used for calibration.

While NMBE and CVRSME may prove useful in describing uncertainty in final savings, it is important to minimize the uncertainty in the simulation inputs. These metrics will not completely capture uncertainty in the inputs.

All software packages acceptable for use in Option D require modelers specify a significant number of physical parameters before simulating a building. Often, many of these parameters have default settings in the software package; however, evaluators can base the parameter inputs on experience or standard practices.

Any parameter not directly based on a physical building or its equipment represents a degree of freedom for calibrating the model against a facility's consumption data.[21] By varying these parameters, the modeler can calibrate the same model to meet uncertainty targets in multiple ways, although for very different reasons.

Lack of a unique calibration point can cause misleading results for NMBE and CVRSME. Furthermore, the resultant calibrations respond differently to changes in other parameters, which can lead to significantly divergent savings estimates. Therefore, it is very important modelers minimize calibration uncertainty *and* they accomplish the calibration for the correct reasons. Modelers should not unreasonably alter inputs simply to reduce NMBE or CVRSME.

The following guidelines minimize uncertainty in the calibration process:

- Experienced simulators (or modelers directly supervised by an experienced simulator must perform the modeling.
- Modelers must document each simulation process step, so reviewers can audit the model, its outputs, and its assumptions.
- Simulators and auditors should determine the most influential default model parameters and confirm their appropriateness.
- Simulated equipment (e.g., HVAC coils, chillers, pumps) should not "auto size" in final simulations.[22]
- Simulators should identify the parameters to which the simulation outputs are most sensitive.[23]

[21] Each parameter must be constrained by a physically realistic range of values.
[22] When specific data are unavailable, auto-sizing can be helpful in determining appropriate coil capacities, fan speeds, etc. However, only use it for initial equipment sizing. Once equipment sizes have been determined, input them directly. Often, modelers must use auto-sizing to define baseline equipment, as the measures impact building loads. In such cases, calculate an *oversize ratio* for as-built equipment and apply it to the baseline simulation.

In addition to quantifying NMBE and CVRSME errors, modelers should analyze the sensitivity of final savings to variations in key model inputs. Modelers should also report such parameters (including their effects on simulated energy savings and the uncertainty in their values) with calibration uncertainty.

[23] Further discussion regarding sensitivity analysis of simulation parameters falls outside this chapter's scope. For additional material on this topic, see Spitler, Fisher, & Zietlow 1989.

5 Sample Design

Use sampling under the following conditions:

- When performing submetering on building equipment
- When performing a detailed survey of an entire building proves impractical.

Evaluators determine the specific targets for sampling certainty and relative precision in the context of the evaluation. For detailed information regarding sample design and for calculating certainty and precision, see the Uniform Method Project's *Chapter 11: Sample Design Cross-Cutting Protocol*.

5.1 Sampling for Submetering

Perform submetering to collect information regarding a building's operational schedules. Monitored systems include lighting, ventilation, large equipment (e.g., data centers), and HVAC zone temperatures. Generally, it is acceptable to assume a coefficient of variation (CV) of 0.5 for most submetering; however, while many of these schedules are a function of the overall building type, significant variation in schedules can occur from space to space within a facility. Therefore, interview site personnel to identify any operational differences (and the magnitude of such differences) within the facility before creating a sample design. Account for variations in operating schedules and usage patterns by using a larger CV or by stratifying unique usage groups. See the Uniform Method Project's *Metering Cross-Cutting Protocols* for additional considerations for commonly monitored equipment.

5.1.1 Example: Monitoring the Lighting Schedule in a Two-Story Office Building

A two-story commercial office building receives a whole-building performance rebate for LEED certification. For the certification process, a DOE2.2 model is built, for which evaluators develop lighting loads and schedules. During the on-site visit, evaluators note the same tenant occupies both floors, and the building remains open from 6:30 a.m. to 10:00 p.m. The evaluators also identify two unique lighting usage patterns:

- Enclosed offices are located on the building's perimeter
- Open office space is located in the building's core.

As the evaluators identified two distinct usage patterns, they should design the sampling to capture the variability within the schedules for both space types.

- As the open office space is located in the building's core, lighting fixtures likely operate continuously during the building's open hours. Additionally, lighting is commonly shared by all workspaces in the building's core. Therefore, a CV of 0.5 is justified and may prove conservative in determining how many fixtures to monitor.
- Lighting fixtures located in enclosed office spaces typically experience significantly more usage variation due to exaggerated behavioral and external influences. Also, the enclosed office space fixtures receive additional light from perimeter windows, thereby reducing the need for interior lighting during daytime hours. These impacts can be exaggerated (or

diminished), depending on fixture control types, building aspects, weather, and times of year. Such additional variability would necessitate a higher assumed CV and additional monitoring points.

5.2 Sampling for Building Surveys

The on-site data collection encompasses a detailed survey of building systems, such as:

- Lighting fixtures
- Plug loads
- HVAC equipment and controls
- Elevator and auxiliary equipment
- Fenestration
- Envelope constructions.

For many buildings, surveyors can perform a complete walk-through and can install monitoring equipment within a single day. However, larger buildings (such as high-rise office buildings, hotel casinos, and hospitals) present logistical and budgetary complexities that make it impractical (and often impossible) to perform a complete facility walk-through. In these cases, it is permissible to perform a walkthrough of a representative sample of building areas and extrapolate the findings to the rest of the building. Evaluators can apply the findings to individual spaces or to entire floors (the exact sample design depends on the facility design, including any considerations, such as access to space).

5.2.1 Example: On-Site Audit of a High-Rise Office Building

A 34-story high-rise commercial building located in a major city's downtown region receives a whole-building performance rebate. Various retail businesses rent the first floor, and various tenants use the remaining floors as office space, including a United States Department of Agriculture office. Evaluators collect data during the on-site visit to build a DOE2.2 model; however, the building owner will only provide evaluation personnel access to the building for a single day.

The building is too large to conduct a thorough walk-through in one day. Additionally, it is expected at least one tenant will have areas within its occupied space that evaluators will not be allowed to access. Therefore, evaluators will have to perform sampling for both floors and space types. Evaluators should audit enough floor space to sufficiently characterize internal loads and usage patterns for each tenant and for the building as a whole. The exact number of floors visited will depend on the number of tenants and on the homogeneity between spaces/floors. The evaluators should:

- Identify unique operating conditions, such as occupancy schedules, lighting power density (and schedules), and equipment power density (and schedules).
- Identify currently vacant areas (or floors).
- Interview facility staff to:

    - Identify differences in space temperatures or ventilation requirements for each tenant
    - Determine variations in building occupancy (by month or as appropriate) since its opening.
- Audit all central plant equipment.
- Sample air distribution system equipment using sampling criteria described in the Uniform Method Project's *Chapter 11: Sample Design Cross-Cutting Protocol.*

6 Program Evaluation Elements

These elements differentiate evaluations of new construction programs from those of other programs:

- Evaluators need significantly more resources to define and justify a hypothetical baseline.
- Evaluators have a limited selection of methods for determining site-level savings.
- Buildings rarely operate at a "steady state" at the time of evaluation.

While this is not a comprehensive list, it specifies critical factors that evaluators must consider in developing an evaluation plan—particularly with regard to budget resources for defining and justifying the baselines used to determine energy savings.

Commonly applied codes (such as ASHRAE 90.1) provide multiple compliance pathways, but leave room for local jurisdictions to maintain their own interpretations. Therefore, evaluators should work with local jurisdictions, program implementers, and evaluation managers and oversight agencies to identify the most appropriate baseline for a building. Further, local jurisdictions may adopt an updated building code during implementation of a program, so the evaluator may have to develop baselines from multiple building codes for a given program year.

Given the limited information available to assess new construction ECMs, using calibrated building simulations is often the only option for determining energy savings. Significant planning ensures:

- Evaluators develop detailed M&V plans each project site
- The evaluation allows sufficient time to perform the analyses.

Evaluators often collect additional information using submetering and/or consumption data analysis. As this information is important for model calibration, the M&V plan should allot sufficient time for a thorough analysis of all submetered data and consumption data.

For programs offering incentives, evaluators usually assess energy efficiency measure performance during the first few years of their operation. During this period, building systems and controls typically require troubleshooting,[24] and buildings have low, but growing, occupancy rates.

Evaluators should also keep in mind that owners (or tenants) may use building spaces differently than as originally designed. Thus, the specific codes or standards governing the originally permitted building drawings may not be appropriate for assessing actual energy use or energy savings. This protocol strongly recommends evaluators consider these and other such factors when calibrating models and simulating annual energy savings.

[24] Troubleshooting is formally done through a commissioning process; however, not all buildings are professionally commissioned. In many facilities, facility management must dial in building controls.

7 References

ASHRAE. (2002). *Guideline 14-2002, Measurement of Energy and Demand Savings.* American Society of Heating, Refrigerating and Air-Conditioning.

ASHRAE. (2007). *Standard 90.1-2007, Energy Standard for Buildings Except Low-Rise Residential.* American Society of Heating, Refrigerating and Air-Conditioning.

CADMAC. (1998). *Appendix J - Quality Assurance Guidelines For Statistical, Engineering, and Self-Report Methods for Estimating DSM Program Impacts.* California DSM Measurement Advisory Council.

CPUC. (2004). *The California Evaluation Framework.* California Public Utilities Commission and the Project Advisory Board.

Crawley, D. B., Griffith, B. T., Hand, J. W., & Kummert, M. (2005). *Contrasting The Capabilities of Building Energy Performance Simulation Programs.*

IPMVP. (2006). *Volume III, Part I: Concepts and Practices for Determining Energy Savings in New Construction.* International Performance Measurement and Verification Protocol.

8 Bibliography

Alereza, T., Besa, A., Cowles, D., & Sebold, F. (1995). "Commercial New Construction Impact Evaluation - A Comprehensive Approach to Analyzing New Construction Incentive Program Results." Prepared for New Construction Programs for Demand-Side Management Conference. ADM Associates Inc.

ASHRAE. (2011). *Standard 189.1-2011, Standard for the Design of High Performance Green Buildings.* American Society of Heating, Refrigerating and Air-Conditioning.

CPUC. (2006). *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals.* California Public Utilities Commission.

Dohrmann, D., Noell, S., & Rittenhouse, B. (1995). "Net Impact Evaluation of A Nonresidential New Construction Program in the Pacific Northwest." Prepared for the New Construction Programs for Demand-Side Management Conference. ADM Associates Inc.

FEMP. (2008). *M&V Guidelines: Measurement and Verification for Federal Energy Projects Version 3.0.* U.S. Department, Energy Federal Energy Management Program. Federal Energy Management Program.

IPMVP. (2001). *Volume I: Concepts and Options for Determining Energy and Water Savings.* International Performance Measurement & Verification Protocol.

Johnson, J., & Nadel, S. (2000). Commercial New Construction Programs: Results from the 90s, Directions for the Next Decade. *Commercial Buildings: Program Design, Implementation, and Evaluation.* American Council for an Energy-Efficient Economy.

Spitler, J., Fisher, D., & Zietlow, D. (1989). *A Primer on the Use of Influence Coefficiencts in Building Simulation.* Simulation '89 Transactions.

Twombly, E. & Hindes, C. (1995). "An Application of ASHRAE 90.1 as a Baseline for Evaluating Commercial New Construction Programs." Prepared for the New Construction Programs for Demand-Side Management Conference. ADM Associates Inc.

Chapter 16: Retrocommissioning Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in September 2014.

Alex Tiessen
Posterity Group
Ottawa, Ontario

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68572
September 2017

Contract No. DE-AC36-08GO28308

Chapter 16: Retrocommissioning Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in September 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in September 2014.

Alex Tiessen
Posterity Group
Ottawa, Ontario

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68572
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter author wishes to thank and acknowledge the following individual for his thoughtful comments and suggestions on drafts of this protocol:

- Greg McGuire of ICF International.

Suggested Citation

Tiessen, A. (2017). *Chapter 16: Retrocommissioning Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68572. http://www.nrel.gov/docs/fy17osti/68572.pdf

Acronyms

BAS	building automation system
HVAC	heating, ventilating, and air conditioning
IPMVP	International Performance Measurement and Verification Protocol
kWh	kilowatt-hour
M&V	measurement and verification
O&M	operation and maintenance
OAT	outdoor air temperature
RCx	retrocommissioning
TMY	typical meteorological year

Table of Contents

List of Figures

List of Tables

1 Measure Description

Retrocommissioning (RCx) is a systematic process for optimizing energy performance in existing buildings. It specifically focuses on improving the control of energy-using equipment (e.g., heating, ventilation, and air conditioning [HVAC] equipment and lighting) and typically does not involve equipment replacement. Field results have shown proper RCx can achieve energy savings ranging from 5% to 20%, with a typical payback of two years or less (Thorne 2003).[1]

The method presented in this protocol provides direction regarding: (1) how to account for each measure's specific characteristics and (2) how to choose the most appropriate savings verification approach.

A study conducted on behalf of Lawrence Berkeley National Laboratory analyzed data from 11 utilities operating RCx programs across the United States. The dataset included 122 RCx projects and more than 950 RCx measures (PECI 2009). Table 1 lists a summary of the most common RCx measures, highlighting the nine measures that represent the majority of the analyzed project savings.

Table 1. Common RCx Measures

RCx Measure	Percentage of Total Savings
Revise control sequence	21%
Reduce equipment runtime	15%
Optimize airside economizer	12%
Add/optimize supply air temperature reset	8%
Add variable frequency drive to pump	6%
Reduce coil leakage	4%
Reduce/reset duct static pressure set point	4%
Add/optimize optimum start/stop	3%
Add/optimize condenser water supply temperature reset	2%

As shown in Table 2 (PECI 2010), RCx measures vary, depending on types of equipment and control mechanisms introduced or optimized. For example, some RCx measures control HVAC equipment according to a predefined schedule, while some measures introduce outdoor air temperature (OAT)-dependent controls.

[1] As discussed in the section "Considering Resource Constraints" of the Introduction chapter to this report, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

Table 2. Categorization of RCx Measures

Control Mechanism	Equipment Type		
	HVAC Airside	HVAC Waterside	Lighting
Scheduled	Matching supply fan schedule to occupancy schedule	Adding/optimizing space setback temperatures	Matching lighting schedule to occupancy schedule
Variable	Optimizing airside economizer	Adding chilled water supply temperature set point reset strategy	Optimizing daylighting control

The classic RCx process helps identify, implement, and maintain improvements to building systems and operations via the following five phases (BPA 2011a).

1. **Planning**. This phase involves screening buildings to determine whether they provide a good fit for RCx by assessing indicators such as equipment age and condition, building energy performance and size, and type of control system. Ideally, facilities should have an existing building automation system (BAS) in good working order, as well as HVAC equipment that is in relatively good condition. A facility without a BAS can install the system; however, the project would then become an HVAC controls and commissioning project rather than an RCx project. When a facility's HVAC equipment nears the end of its useful life, undertaking RCx may not be appropriate because control measures could become obsolete with replaced equipment.
2. **Investigation**. The investigation phase involves analyzing facility performance by reviewing building documentation; performing diagnostic monitoring and functional tests; interviewing staff; identifying a list of recommended improvements; and estimating savings and costs. Evaluators should clearly differentiate valid RCx measures that meet program eligibility guidelines from retrofit measures and/or operation and maintenance (O&M) activities at this phase.
3. **Implementation**. The implementation phase involves prioritizing recommended measures and developing an implementation plan; implementing the measures; and testing to ensure proper operation. Implementation often entails an iterative approach, as the evaluator may need to determine the final control set points through several stages of modification and assessment. These stages ensure building equipment continues to operate properly and maintains the occupants' comfort. Typically, evaluators will review a facility's BAS to assess how effectively RCx measures operate.
4. **Turnover**. The turnover phase involves updating building documentation (e.g., system operation manuals); developing and presenting a final report; and training building operators on proper O&M.
5. **Persistence**. The persistence phase involves monitoring and tracking energy use over time; continually implementing persistence strategies (e.g., refining control measures or enhancing O&M procedures) to sustain savings; and documenting ongoing changes. Depending on the availability of resources and the timeline, program stakeholders may not always actively support this phase.

2 Application Conditions of Protocol

The RCx program design includes activities intended to overcome a number of market barriers, as listed in Table 3.

Table 3. RCx Market Barriers

Market Segment	Barrier	Opportunities
Supply-Side Actors, End Users	No tangible examples of RCx performance *in situ*	Undertaking pilot projects
Supply-Side Actors	Lack of service provider capacity for undertaking the RCx investigation and implementation phases	Training for service providers
End Users	Lack of awareness and understanding of the RCx benefits	Education to increase building owner and operator awareness
End Users	Cost of undertaking RCx	Incentives

Ideally, energy-efficiency programs overcome these barriers through various activities that address available opportunities. Retrocommissioning programs may include some or all of the following activities:

- **Pilot projects**. Program administrators sometimes fund pilot projects to demonstrate the benefits of RCx to end users in their target markets. Evaluators can verify pilot savings using the methods presented later in this protocol and, in theory, these savings will attract participants to the program.
- **Training**. Program administrators sometimes fund or develop training for service providers. In some jurisdictions, service providers do not routinely provide RCx services to their customer base. Thus, to develop RCx capacity in the market, program administrators might offer training to service providers on how to provide common practice RCx investigation and implementation services. Service providers may also require training on how to sell these services to their clients.
- **Education**. Program administrators sometimes develop educational materials and hold events or workshops for end users. Prior to making a decision to undertake RCx activities in their facilities, building management and building operators need to understand the business case for RCx. Detailed case studies showcasing project savings are an example of education tools program staff can use to facilitate this decision-making process.
- **Incentives**. Program administrators often provide incentives to undertake the RCx investigation, implementation, and persistence phases. Even though the payback for RCx measures is typically low, end users often require incentives to encourage them to move forward with projects.[2] Incentives may also encourage end users to undertake projects sooner—or with a greater scope—than they would have without market intervention.

[2] Some programs may impose a penalty rather than an incentive. For example, if participants fail to implement the measures that fell below a certain payback threshold identified during the investigation phase, they may not be eligible for the full investigation phase incentive.

This protocol provides structured methods for determining energy savings resulting from the implementation of RCx measures. The approaches described here provide direction on how to verify savings consistently from pilot projects, as well as from projects implemented by program participants. It does not address savings achieved through training or through market transformation activities.

3 Savings Calculations

Specific savings calculations[3] for RCx measures inherently vary, due to the breadth of possible RCx measures, which can differ by type of equipment or control mechanism. This section presents a high-level gross energy savings equation that is applicable to all RCx measures. Section 4, *Measurement and Verification Plan*, includes detailed directions for calculating savings for specific measure categories.

Use the following general equation (EVO 2012) to determine energy savings:

Equation 1

Energy Savings = (Baseline Energy – Reporting Period Energy) ± Routine Adjustments ± Nonroutine Adjustments

Where,

Energy Savings	= First-year energy consumption savings.
Baseline Energy	= Preimplementation consumption.
Reporting Period Energy	= Postimplementation consumption.
Routine Adjustments	= Adjustments made to account for routinely changing independent variables (variables that drive energy consumption). If applicable, normalize savings to typical meteorological year (TMY[4]) weather data, as well as other significant independent variables (e.g., occupancy, production data).
Nonroutine Adjustments	= Adjustments made to account for parameters typically not expected to change during the implementation period. Account for these parameters if they change *and* this change influences the reporting period energy use (e.g., changes to a facility's building envelope during implementation of an RCx HVAC measure). Evaluators only need to consider nonroutine adjustments if verifying savings using Option C of the International Performance Measurement and Verification Protocol (IPMVP).[5]

Determining RCx demand savings is not a straightforward extension of verified consumption savings (unlike lighting retrofits, where evaluators can easily apply established load savings

[3] As presented in the Introduction, the protocols focus on gross energy savings and do not include other parameter assessments, such as net-to-gross, peak coincidence factors, or cost-effectiveness.

[4] Evaluators should use the most recent typical meteorological year dataset. As of January 2014, the most comprehensive national typical meteorological year dataset is TMY3. Evaluators should confer with the local jurisdiction to see if they should use a different regional dataset.

[5] Option C is the "whole-facility approach" to verifying savings.

profiles to consumption savings data). For RCx projects, load savings profiles vary depending on the type of measures implemented and the distribution of these measures. If applicable, evaluators should produce load savings profiles on a measure-by-measure basis,[6] aggregate these profiles, and then apply site-specific coincidence factors to determine coincident peak demand savings at the project level.

[6] Alternatively, if verifying savings by following Option C or D of the IPMVP, evaluators can measure or compute aggregate project-level load savings profile and negate the requirement to build up the profile on a measure-by-measure basis. If using Option C, evaluators should investigate whether data from advanced metering infrastructure (e.g., interval meters) is available to increase the accuracy of billing data analyses.

4 Measurement and Verification Plan

This section outlines the recommended approaches to determining RCx energy savings and provides directions on how to use the approaches under the following headings:

- Measurement and verification (M&V) method
- Data collection
- Interactive effects
- Specific savings equations
- Regression model direction
- Deemed spreadsheet tool functionality requirements.

4.1 Measurement and Verification Method

There is a structured method for determining the most appropriate approach to verifying RCx energy savings. This method balances the need for accurate energy-savings estimates with the need to keep M&V costs in check, relative to project costs and anticipated energy savings. Depending on which measures are implemented, different approaches to estimating the savings are appropriate. Following the IPMVP, the options are:

- Option A—Retrofit Isolation: Key Parameter Measurement
- Option B—Retrofit Isolation: All Parameter Measurement
- Option C—Whole Facility
- Option D—Calibrated Simulation

Measurement is inherent with most RCx projects because RCx measures typically involve modifications made through a facility's BAS. As mentioned, RCx implementation (an iterative process) often leverages metered data to evaluate and optimize changes throughout the process. Therefore, in many cases, a retrofit isolation approach adhering to Option A or Option B of the IPMVP proves most logical. That said, scenarios exist where Option C, Option D, or even a deemed approach may be more appropriate. Figure 1 presents a decision flow chart for determining the approaches to follow.

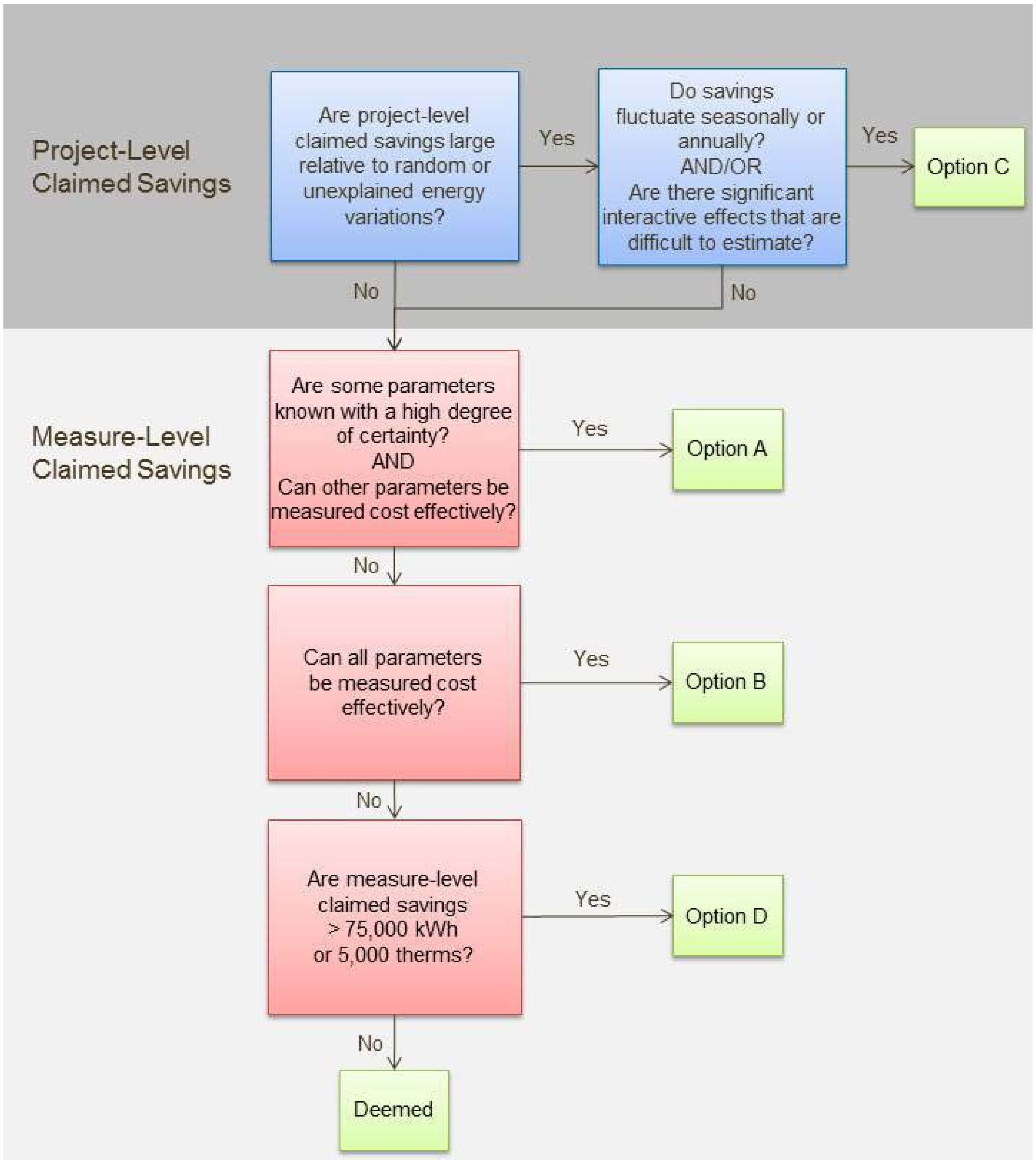


Figure 1. RCx approach—decision flow chart

The decision flow chart accounts for factors such as the magnitude of estimated savings and the measurement's cost-effectiveness. Begin the process by considering project-level savings:

- **Option C.** Use a whole-facility approach—adhering with Option C of the IPMVP—if estimated project-level savings are large compared to the random or unexplained energy variations that occur at the whole-facility level[7] and if savings fluctuate over a seasonal or annual cycle (e.g., savings that fluctuate depending on OAT). This approach is likely the most cost-effective approach for verifying savings. The whole-facility approach is

[7] Typically savings should exceed 10% of the baseline energy for a particular meter (e.g., electricity meter) to confidently discriminate the savings from the baseline data when the reporting period is shorter than two years (EVO 2012).

relatively inexpensive because evaluators can use utility billing data for the analysis. The downside of the approach is that evaluators cannot perform verification until after collecting a full season or year of reporting period data and monitoring and documenting any changes to the facility's static factors[8] over the course of the measurement period. Even if savings remain consistent month to month, Option C may provide the best approach if project measures cause complex, significant interactive effects. Such interactive effects are, by nature, difficult to estimate accurately. Also, if the effects are significant (large, relative to direct-measure savings), evaluators will be required to use a whole-facility approach to measure impacts accurately. The reduced heating and cooling energy resulting from schedule changes to an air-handling unit, when control modifications have also been undertaken for both the heating and cooling systems, is an example of a complex significant interactive effect warranting Option C.

If Option C is ruled out, consider performing verification on a measure-by-measure basis:

- **Option A.** If measures involve some parameters known with a high degree of certainty *and* other parameters can be measured cost-effectively, use a retrofit isolation approach adhering to Option A of the IPMVP. In many cases, evaluators can collect metered data directly from the facility's BAS. If required, the facility can add control points to the BAS, either as part of the implementation process or specifically for M&V purposes. Where the BAS cannot provide the information, use temporary meters to collect data (provided that costs are not prohibitive).
- **Option B.** If a given measure's parameters are uncertain but can be measured cost-effectively, use a retrofit isolation approach, adhering to Option B of the IPMVP. Again, collect metered data (similar to Option A) either through the BAS or by using temporary meters.
- **Option D.** For measures where it is prohibitive to meter all required parameters, use a calibrated simulation approach adhering to Option D of the IPMVP. Undertake calibrations in two ways: (1) calibrate the simulation to the actual baseline or reporting consumption data and (2) confirm the reporting period inputs via the BAS front-end system, when possible.[9,10]
- **Deemed.** Finally, if a measure is relatively common[11] and its estimated savings are small, evaluators can deem savings rather than simulate them. Use this approach for common measures with savings less than 75,000 kilowatt-hours (kWhs) or 5,000 therms[12] (PECI

[8] Many factors can affect a facility's energy consumption, even though evaluators do not expect them to change. These factors are known as "static factors" and include the complete collection of facility parameters that are generally expected to remain constant between the baseline and reporting periods. Examples include: building envelope insulation, space use within a facility, and facility square footage.
[9] In many cases, the simulation should represent the entire facility; however, in some cases, depending on the facility's wiring structure, a similar approach could be applied to building submeters, such as distribution panels that include the affected systems.
[10] See the Uniform Method Project's *Commercial New Construction Protocol* for more information on using Option D.
[11] If regulators are involved, going through the effort of deeming savings for a rare measure can be burdensome.
[12] Program administrators and evaluators may wish to customize these thresholds for particular programs and/or jurisdictions.

2010). Use a spreadsheet tool to calculate savings, adhering to functionality requirements presented later in the protocol.

4.2 Data Collection

Depending on the approach followed, these M&V elements will require particular consideration:

- The measurement boundary
- The measurement period and frequency
- The functionality of measurement equipment being used
- The savings uncertainty.

4.2.1 Measurement Boundary

For measures evaluators assess using Option A or Option B and that require metering external to the BAS, it will be important to define the measurement boundary. When determining boundaries—the location and number of measurement points required—consider the project's complexity and expected savings:

- While a narrow boundary simplifies data measurement (e.g., a single piece of equipment), variables driving energy use outside the boundary (i.e., interactive effects) still need to be considered.
- A wide boundary will minimize interactive effects and increase accuracy (e.g., systems of equipment like chilled water plants and air-handling units). However, as M&V costs may also increase, it is important to ensure the expected project savings justify the increased M&V costs.

4.2.2 Measurement Period and Frequency

For all measures assessed with Option A or Option B, consider two important timing metrics:

- The measurement period (the length of the baseline and reporting periods)
- The measurement frequency (how regularly to take measurements during the measurement period).

As a general rule, choose the measurement period to capture a full cycle of each operating mode. For example, if there is a control modification to heating equipment, collect data over the winter and shoulder seasons.

Choose the measurement frequency by assessing the type of load measured:

- Spot measurement: For constant loads, measure power briefly, preferably over two or more intervals.
- Short-term measurement: For loads predictably influenced by independent variables (e.g., HVAC equipment influenced by OAT), take short-term consumption measurements over the fullest range of possible independent variable conditions, given M&V project cost

and time limitations. 13 For systems expected to have nonlinear dependence (such as air-handling units with outside air economizers), measurements should incorporate sufficient range to characterize the full breadth of conditions.

- Continuous measurement: For variable loads, measure consumption data continuously, or at appropriate discrete intervals, over the entire measurement period.

See Section 4.4, *Specific Saving Equations*, for direction regarding measurement periods and frequency for specific measure types.

4.2.3 Measurement Equipment

When meters external to the BAS are required, follow these guidelines to select a meter:[14]

- Size the meter for the range of values expected most of the time.
- Select the meter repeatability and accuracy that fits the budget and intended use of the data.
- Install the meter as recommended by the manufacturer.
- Calibrate the meter before it goes into the field, and maintain calibration as recommended by the manufacturer. If possible, select a meter with a recommended calibration interval that is longer than the anticipated measurement period.

If BAS data is used, evaluators should exercise due diligence by determining when the BAS was last calibrated and by checking the accuracy of the BAS measurement points.

4.2.4 Savings Uncertainty

If possible, quantify the accuracy of measured data[15] and, if practical, conduct an error propagation analysis to determine overall impacts on the savings estimate.

4.3 Interactive Effects

For projects following Option A, Option B, or deemed approaches, consider and estimate interactive effects if they are significant. For example, if a facility reduces an air-handling unit supply fan schedule, not only will direct fan savings be achieved, but significant cooling and heating energy savings may be realized due to decreases in conditioned ventilation air supplied to the space.

Estimate interactive effects using equations that apply the appropriate engineering principles. Ideally, use a spreadsheet tool adhering to the same functionality requirements discussed in

---

[13] For example, if a chiller plant undergoes control modifications, the measurement frequency should be long enough to capture the full OAT operating range. In a temperate climate zone, evaluators can accomplish this by taking measurements over a four-week period in the shoulder season and another four-week period during the summer season.

[14] For more information on selecting measurement equipment, see the Uniform Methods Project's *Metering Cross-Cutting Protocols*.

[15] Metering accuracy is only one element of savings uncertainty. Inaccuracies also result from modeling, sampling, interactive effects, estimated parameters, data loss, and measurements being taken outside of a meter's intended range.

Section 4.6 for the deemed spreadsheet tool to conduct these analyses. When interactive effects are large, it may be possible to measure them rather than apply engineering estimates. In the "supply fan" example discussed in the paragraph above, an evaluator can meter the chilled water plant to determine the cooling load reduction.

Interactive effects for projects being verified using Option C or Option D are typically included in facility-level savings estimates.

4.4 Specific Savings Equations

If following Option A or Option B, verify savings using equations matching a given measure's characteristics—specifically, whether savings are dependent on independent variables (such as OAT) and the control mechanism for affected equipment.

Figure 2 shows the three categories of savings equations, with further explanations following the flow chart.

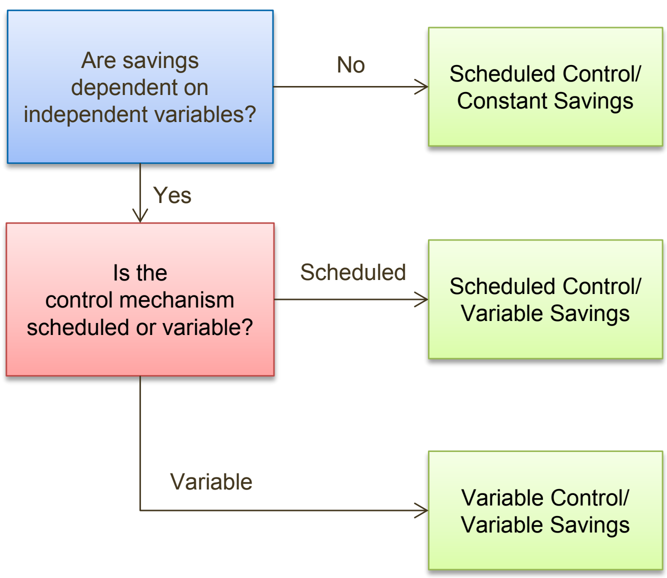


Figure 2. Savings equation categories

4.4.1 Scheduled Control/Constant Savings

This savings equation category encompasses scheduled control measures on equipment not influenced by independent variables (such as OAT); therefore, this is the most straightforward equation category.

Lighting schedule optimization is an example of a measure verified using this savings equation category. In this example, lighting is turned off according to a schedule (scheduled control), and constant savings is achieved while it is off (constant savings).[16]

[16] While a single piece of equipment (one lighting fixture) may have a constant load, the system (lighting throughout a building) may have some variability. In a lighting system that includes a degree of occupant control (such as

Equation 2

Scheduled Control/Constant Savings = Baseline Energy – Reporting Period Energy

Where,

Scheduled Control/Constant Savings	= First-year energy consumption savings resulting from a scheduled control measure with constant savings.
Baseline Energy	= $HRS_{baseline}$ x $kW_{controlled}$
Reporting Period Energy	= $HRS_{reporting}$ x $kW_{controlled}$

And,

$HRS_{baseline}$	= Annual operating hours during the baseline: if this parameter is not known with a high degree of certainty, take short-term measurements for the duration of each existing schedule type.
$HRS_{reporting}$	= Annual operating hours during the reporting period: take short-term measurements for the duration of each new schedule type.
$kW_{controlled}$	= Electric demand controlled by scheduling measure: if this parameter is not known with a high degree of certainty, take spot measurements during the baseline or reporting period.

4.4.2 Scheduled Control/Variable Savings

This savings equation category encompasses scheduled control measures on equipment influenced by independent variables (such as OAT). Space setback temperature optimization provides an example of a measure verified using this savings equation category. In this example, the heating space temperature set point is lowered according to a schedule during unoccupied hours (scheduled control), and the savings achieved will vary, depending on OAT (variable savings).

Following Equation 3, Table 4 lists the five-step process for determining adjusted baseline and reporting period energy consumption.

---

switches in private offices) nearly 100% of fixtures may operate midday, but substantially fewer may be on at the beginning or end of the day when the savings due to scheduling would likely occur.

Equation 3

Scheduled Control/Variable Savings = Adjusted Baseline Energy – Adjusted Reporting Period Energy

Where,

Scheduled Control/Variable Savings = First-year energy consumption savings resulting from a scheduled control measure with variable savings.

Adjusted Baseline Energy = $\sum_{All\ Schedule\ Types} Adj\ Baseline\ Consumption_{Schedule\ Type}$ and determined through the five-step process listed in Table 5.

Adjusted Reporting Period Energy = $\sum_{All\ Schedule\ Types} Adj\ Reporting\ Period\ Consumption_{Sche}$ determined through the five-step process listed in Table 5.

Table 4. Adjusted Consumption for Scheduled Control/Variable Savings Measures

<table>
<tr><th>Step</th><th colspan="3">Details</th></tr>
<tr><td>Develop baseline/reporting regression model(s) by measuring equipment operation and independent variables.</td><td colspan="3">Take short-term measurements at representative load levels for the affected equipment for each schedule type.

Take coincident measurements of the independent variable(s).

Do a regression analysis to determine the relationship between independent variables and equipment load. This relationship should be expressed in terms of an equation (baseline/reporting period model).

Note: if there are schedules for occupied and unoccupied times during the reporting period, evaluators will need two regression models, one for each set of data.</td></tr>
<tr><td rowspan="3">Develop a bin operating profile[a] by normalized independent variable data.</td><td colspan="3">Develop bin data tables presenting the following data (one table for each schedule type):</td></tr>
<tr><td>Independent Variable</td><td>Load</td><td>Annual Hours</td></tr>
<tr><td>Create approximately 10 bins over the normalized independent variable data range (if the equipment’s energy consumption varies depending on weather, use TMY data).</td><td>Calculate the normalized load by applying the baseline/reporting period regression model to the midpoint of each bin.</td><td>Use short-term measured data to estimate hours of operation within each bin or base this on TMY data and the equipment operating schedule.</td></tr>
<tr><td>Calculate the baseline/reporting period consumption at each load bin for each schedule type.</td><td colspan="3">$\text{Adjusted Consumption}_{\text{Load,Schedule Type}} = \text{Load}_{\text{Schedule Type}} \times \text{Annual Hrs}_{\text{Schedule Type}}$</td></tr>
<tr><td>Sum the consumption savings across bins for each schedule type.</td><td colspan="3">$\sum_{All\ Load\ Bins_{Schedule\ Type}} Adj\ Consumption_{Load,Schedule\ Type}$</td></tr>
<tr><td>Sum the consumption savings across schedule types.</td><td colspan="3">$\sum_{All\ Schedule\ Types} Adj\ Consumption_{Schedule\ Type}$</td></tr>
</table>

[a] Alternatively, if the independent variable is OAT, evaluators can develop an hourly profile over the full operating schedule of the affected equipment.

4.4.3 Variable Control/Variable Savings

This savings equation category encompasses variable control measures on equipment influenced by independent variables, such as OAT. Introducing a chilled water supply temperature set point reset strategy serves as an example of a measure verified through this savings equation category. In this example, the chilled water supply temperature set point is determined depending on OAT (variable control), and the savings achieved will vary depending on OAT (variable savings).

Following Equation 4, Table 5 lists the four-step process for determining the adjusted baseline and reporting period energy consumption.

Equation 4

Variable Control/Variable Savings = Adjusted Baseline Energy – Adjusted Reporting Period Energy

Where,

Variable Control/Variable Savings = First-year energy consumption savings resulting from a variable control measure with variable savings.

Adjusted Baseline Energy = $\sum_{All\ Load\ Bins} Adj\ Baseline\ Consumption_{Load}$ determined through the four-step process listed in Table 6.

Adjusted Reporting Period Energy = $\sum_{All\ Load\ Bins} Adj\ Reporting\ Period\ Consumption_{Load}$ determined through the four-step process listed in Table 6.

Table 5. Adjusted Consumption for Variable Control/Variable Savings Measures

<table>
<tr><th>Step</th><th colspan="3">Details</th></tr>
<tr><td>Develop baseline/ reporting regression model(s) by measuring equipment operation and independent variables.</td><td colspan="3">Take short-term measurements at representative load levels for the affected equipment for each schedule type.

Take coincident measurements of the independent variable(s).

Do a regression analysis to determine the relationship between independent variables and equipment load. This relationship should be expressed in terms of an equation (baseline/reporting period model).</td></tr>
<tr><td rowspan="3">Develop a bin operating profile[a] by normalized independent variable data.</td><td colspan="3">Develop bin data tables presenting the following data:</td></tr>
<tr><td>Independent Variable</td><td>Load</td><td>Annual Hours</td></tr>
<tr><td>Create approximately 10 bins over the normalized independent variable data range (e.g., if the equipment's energy consumption varies depending on weather, use TMY data).</td><td>Calculate the normalized load by applying the baseline/reporting period regression model to the midpoint of each bin.</td><td>Use short-term measured data to estimate hours of operation within each bin, or base this on TMY data and the equipment operating schedule.</td></tr>
<tr><td>Calculate the baseline/reporting period consumption at each load bin.</td><td colspan="3">Adjust Consumption
Adj $Consumption_{Load}$ = Load x Annual Hours</td></tr>
<tr><td>Sum the consumption savings across bins.</td><td colspan="3">$\sum_{All\ Load\ Bins} Adj\ Consumption_{Load}$</td></tr>
</table>

[a] Alternatively, if the independent variable is OAT, evaluators can develop an hourly profile over the full operating schedule of the affected equipment.

4.5 Regression Modeling Direction

Calculating normalized savings for the majority of projects—whether following the IPMVP's Option A, Option B, or Option C— will require the development of a baseline and reporting period regression model.[17] Use one of the following three types of analysis methods to create the model:

- *Linear Regression:* For one routinely varying significant parameter (e.g., OAT).[18]
- *Multivariable Linear Regression*: For more than one routinely varying significant parameter (e.g., OAT, occupancy).
- *Advanced Regression*: For a multivariable, nonlinear fit requiring a polynomial or exponential model.[19]

Develop all models in accordance with best practices and only use them when they are statistically valid (see Subsection 4.5.2, *Testing Model Validity*). If no significant independent variables arise (as with a lighting schedule measure), evaluators are not required to use a model because calculated savings will be inherently normalized.

4.5.1 Recommended Methods for Model Development

Use energy and independent variable data that is representative of a full cycle of operation. For example, if facility staff implement a heating space temperature setback measure, collect energy data across the full range of OAT for each of the operating schedules (occupied and unoccupied) for each season, as shown in Table 6.

Table 6. Example of Data Required for Model Development

	Shoulder Season	Winter Season
Occupied Hours	Short-term energy measurements during occupied hours. Measurements should be representative of the full range of shoulder-season OAT (approximately 10 OAT bins).	Short-term energy measurements during occupied hours. Measurements should be representative of the full range of winter-season OAT (approximately 10 OAT bins).
Unoccupied Hours	Short-term energy measurements during unoccupied hours. Measurements should be representative of the full range of shoulder-season OAT (approximately 10 OAT bins).	Short-term energy measurements during unoccupied hours. Measurements should be representative of the full range of winter-season OAT (approximately 10 OAT bins).

[17] This could either be a single regression model that uses a dummy variable to differentiate the baseline/reporting period data or two independent models for the baseline and reporting period, respectively.
[18] One of the most common linear regression models is the three-parameter change point model. For example, a model that represents cooling electricity consumption will have one regression coefficient that describes nonweather-dependent electricity use, a second regression coefficient that describes the rate of increase of electricity use with increasing temperature, and a third parameter that describes the change point temperature, also known as the balance point temperature, where weather-dependent electricity use begins.
[19] Evaluators may need to use advanced regression methods if RCx activities impact manufacturing or industrial process equipment.

Analyze the data collected to identify outliers. Only remove outliers when there is a tangible explanation to support the erratic data points. Discussion of how to identify outliers is outside the scope of this protocol.

4.5.2 Testing Model Validity

To assess the model's accuracy, begin by reviewing the parameters in Table 7 (EVO 2012).

Table 7. Model Statistical Validity Guide

Parameter Evaluated	Description	Suggested Acceptable Values
Coefficient of determination (R^2)	A measure of the extent that the regression model explains variations in the dependent variable from its mean value.	> 0.75
T-statistic (absolute value)	An indication of whether regression model coefficients are statistically significant.	> 2[a]
Mean bias error	An indication of whether the regression model overstates or understates actual energy consumption.	Will depend on the measure, but generally: < ±5%

[a] Determine the t-statistic threshold based on the evaluator's chosen confidence level; a 95% confidence level requires a t-statistic of 1.96. Evaluators should determine an acceptable confidence level depending on project risk (i.e., savings risk), budget, and other considerations.

A model outside the suggested range indicates parameter coefficients that are relatively poorly determined, with the result that normalized consumption will have relatively high statistical prediction error. Ordinarily, evaluators should not use such a model for normalization, unless the analysis includes appropriate statistical treatment of this prediction error. Discussion of how to proceed in such circumstances is outside the scope of this protocol.

When possible, attempt to enhance the regression model by:

- Increasing or shifting the measurement period
- Incorporating more data points
- Including independent variables previously unidentified
- Eliminating statistically insignificant independent variables.

Also, when assessing model validity, consider coefficient of variation of the root mean squared error, fractional savings uncertainty, and residual plots. Refer to ASHRAE Guideline 14-2002 and Bonneville Power Administration's *Regression for M&V: Reference Guide* for direction on how to assess these additional parameters.

4.6 Deemed Spreadsheet Tool Functionality Requirements

When collecting measured energy data is not cost-effective and claimed (*ex ante)* savings estimates for a given measure are sufficiently small (75,000 kWh or 5,000 therms), use a deemed approach to calculate savings. In this scenario, the protocol recommends using a spreadsheet tool to calculate savings, and this tool should meet these general requirements:

- **Ensure model transparency.** A third party should be able to review the spreadsheet tool and clearly understand how the evaluator derived all savings outputs. To this end, clearly explain and reference all inputs and calculation algorithms within the spreadsheet. Do not lock or hide cells or sheets and check to ensure all links work properly.
- **Use relevant secondary data.** When using secondary data as inputs to savings algorithms, ensure they are relevant to the project's region or jurisdiction. Substantiate input relevancy within the spreadsheet. For example, if using assumed values for hours of operation for heating equipment, take these secondary data from a regional resource (e.g., a technical resource manual from the most applicable demand-side management authority).
- **Verify input elements—either on site or through the BAS front-end system.** Even when using a deemed approach, verify and update some inputs with actual site observations (rather than solely relying on secondary data). For example, confirm a new lighting schedule through the BAS front-end system and note it in the spreadsheet tool.
- **Establish default values for unverifiable parameters.** Use default values for parameters that cannot be verified. For example, clearly state assumed values for motor efficiencies and load factors.

The Building Optimization Analysis Tool,20 developed by Portland Energy Conservation Inc., (PECI 2010) provides an example of benchmark for RCx spreadsheet tools. Although the protocol does not require the following level of rigor, ideally, a best-practice spreadsheet tool should:

- Incorporate regional TMY data.
- Incorporate regional building archetype templates.
- Undergo a calibration process by using measured data from previous regional projects to test algorithms.

[20] Download the tool for free at: www.cacx.org/resources/rcxtools/spreadsheet_tools.html.

5 Sample Design

Consult the Uniform Methods Project's *Sample Design Cross-Cutting Protocols* for general sampling procedures if the RCx program project population is sufficiently large or if the evaluation budget is constrained. Ideally, use stratified sampling to partition RCx projects by measure type, facility type, and/or project size. Stratification ensures evaluators can confidently extrapolate sample findings to the remaining project population. Regulatory or program administrator specifications typically govern the confidence and precision-level targets that influence sample size.

6 Other Evaluation Issues

When claiming lifetime and net program RCx impacts, evaluators should consider persistence and net-to-gross in addition to first-year gross impact findings.

6.1 Persistence

Persistence of savings encompasses both the retention and the performance degradation of measures. Evaluators should consider persistence on a program-by-program basis because the persistence of RCx projects can vary widely depending on the distribution of measure types implemented and, perhaps more significantly, on how well facility staff maintains the modifications. Consult the Uniform Methods Project's *Assessing Persistence and Other Evaluation Issues Cross-Cutting Protocols* for more information.

6.2 Net to Gross

Consult the Uniform Methods Project's *Estimating Net Energy Savings: Common Practices* for a discussion about determining net program impacts at a general level, including direction on how to assess freeridership. Supplementary to that chapter, however, evaluators may consider assessing participant spillover if evidence emerges of participants implementing no-cost measures. This would specifically apply to no-cost measures identified during the investigation phase, but not explicitly included under the scope of program-funded RCx implementation activities.

If no-cost measures exist and there are no savings claims, the attribution evaluation may involve interviews with building operators and their service providers to obtain estimates of the savings magnitude resulting from these measures. Participant spillover would positively influence the program's overall net-to-gross factor.

7 References

BPA. (2011a). *Existing Building Commissioning: An M&V Protocol Application Guide.* Bonneville Power Administration.

EVO. (2012). *International Performance Measurement and Verification Protocol—Concepts and Options for Determining Energy and Water Savings, Volume 1.* Efficiency Valuation Organization.

Itron Inc. (2008). *San Diego Gas and Electric 2005–2005 Retrocommissioning Program Final Report*. Prepared for Portland Energy Conservation Inc.

Luskay, L., Haasl, T., Irvine, L., and Frey, D. (2002). *Retrocommissioning Case Study—Applying Building Selection Criteria for Maximum Results*. Prepared for the International Conference for Enhanced Building Operations.

PECI. (2009). *A Study on Energy Savings and Measure Cost Effectiveness of Existing Building Commissioning*. Prepared for Lawrence Berkeley National Laboratory. Portland Energy Conservation Inc.

PECI. (2010). *Building Optimization Analysis Tool—Final Project Report to PG&E*. Prepared for Pacific Gas and Electric. Portland Energy Conservation Inc.

Thorne, J., and Nadel, S. (2003). *Retrocommissioning: Program Strategies to Capture Energy Savings in Existing Buildings*. Prepared for American Council for an Energy-Efficient Economy.

Tso, B., Hall, N., Lai, P., and Pulliam, R. (2007). *How Much Does Retrocommissioning Really Save? Results from Three Commissioning Program Evaluations in California.* Prepared for the International Energy Program Evaluation Conference.

8 Bibliography

ASHRAE. (2002). *Guideline 14-2002: Measurement of Energy and Demand Savings*. American Society of Heating, Refrigerating and Air-Conditioning.

BPA. (2011b). *Regression for M&V: Reference Guide.* Bonneville Power Administration.

California Commissioning Collaborative. (2006). *California Commissioning Guide: Existing Buildings*.

California Public Utilities Commission. (2006). *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*.

Dethman, L., Kunkle, R., and Degens, P. (2007). *Chasing O&M Savings: Process Lessons from Two Pilot Programs in the Northwest*. Prepared for the International Energy Program Evaluation Conference.

Natural Resources Canada. (2009). *Recommissioning (RCx) Tools Assessment Report.*

Northeast Energy Efficiency Partnerships. (2010). *Regional EM&V Methods and Savings Assumptions Guidelines.*

RLW Analytics, Inc. (2008). *2004–2005 Los Angeles County—Internal Services Department/Southern California Edison/Southern California Gas Company Energy Efficiency Partnership Impact Evaluation Study*. Prepared for California Public Utilities Commission.

U.S. Department of Energy Federal Energy Management Program. (2008). *M&V Guidelines: Measurement and Verification for Federal Energy Projects—Version 3.0*.

U.S. Environmental Protection Agency. (2008). *ENERGY STAR Building Upgrade Manual*.

Chapter 17: Residential Behavior Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – December 2017

This version supersedes the version originally published in January 2015. The content in this version has been updated.

James Stewart
Cadmus
Waltham, Massachusetts

Annika Todd
Lawrence Berkeley National Laboratory
Berkeley, California

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68573
October 2017

Contract No. DE-AC36-08GO28308

Chapter 17: Residential Behavior Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – December 2017

This version supersedes the version originally published in January 2015. The content in this version has been updated.

James Stewart
Cadmus
Waltham, Massachusetts

Annika Todd
Lawrence Berkeley National Laboratory
Berkeley, California

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68573
October 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

These methods, processes, or best practices ("Practices") are provided by the National Renewable Energy Laboratory ("NREL"), which is operated by the Alliance for Sustainable Energy LLC ("Alliance") for the U.S. Department of Energy (the "DOE").

It is recognized that disclosure of these Practices is provided under the following conditions and warnings: (1) these Practices have been prepared for reference purposes only; (2) these Practices consist of or are based on estimates or assumptions made on a best-efforts basis, based upon present expectations; and (3) these Practices were prepared with existing information and are subject to change without notice.

The user understands that DOE/NREL/ALLIANCE are not obligated to provide the user with any support, consulting, training or assistance of any kind with regard to the use of the Practices or to provide the user with any updates, revisions or new versions thereof. DOE, NREL, and ALLIANCE do not guarantee or endorse any results generated by use of the Practices, and user is entirely responsible for the results and any reliance on the results or the Practices in general.

USER AGREES TO INDEMNIFY DOE/NREL/ALLIANCE AND ITS SUBSIDIARIES, AFFILIATES, OFFICERS, AGENTS, AND EMPLOYEES AGAINST ANY CLAIM OR DEMAND, INCLUDING REASONABLE ATTORNEYS' FEES, RELATED TO USER'S USE OF THE PRACTICES. THE PRACTICES ARE PROVIDED BY DOE/NREL/ALLIANCE "AS IS," AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL DOE/NREL/ALLIANCE BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER, INCLUDING BUT NOT LIMITED TO CLAIMS ASSOCIATED WITH THE LOSS OF PROFITS, THAT MAY RESULT FROM AN ACTION IN CONTRACT, NEGLIGENCE OR OTHER TORTIOUS CLAIM THAT ARISES OUT OF OR IN CONNECTION WITH THE ACCESS, USE OR PERFORMANCE OF THE PRACTICES.

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter authors wish to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Ingo Bensch of Evergreen Economics
- Debbie Brannan, Bill Provencher, Kevin Cooney, Carly Olig, and Frank Stern of Navigant
- Cheryl Jenkins of Vermont Energy Investment Corporation
- M. Sami Khawaja of Cadmus
- Maggie McCarey, Marisa Uchin, and Alessandro Orfei of Oracle Utilities (Opower)
- Tim Guiterman and John Backus Mayes of Energy Savvy
- Julie Michals of 4TheFuture.

Suggested Citation

Stewart, J.; Todd, A. (2017). *Chapter 17: Residential Behavior Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68573. http://www.nrel.gov/docs/fy17osti/68573.pdf

Acronyms

BB	behavior-based
DiD	difference-in-differences
IPMVP	International Performance Measurement and Verification Protocol
ITT	intent-to-treat
IV	instrumental variable
LATE	local average treatment effect
OLS	ordinary least squares
PG&E	Pacific Gas & Electric
RCT	randomized control trial
RED	randomized encouragement design
SEE Action	State and Local Energy Efficiency Action
TOT	treatment effect on the treated
UMP	Uniform Methods Project

Protocol Updates

The original version of this protocol was published in January 2015. The authors updated the protocol by making the following changes:

- Incorporated findings from recent research comparing the accuracy of savings estimates from randomized experiments and quasi-experiments
- Presented new developments in the estimation of energy savings from behavior-based programs, including the post-period only model with pre-period controls (Allcott 2014)
- Updated the discussion of randomized encouragement designs to emphasize the importance of having large sample sizes or a sufficient proportion of compliers as well as the application of instrumental variables two-stage least squares for obtaining estimates of the local average treatment effect
- Incorporated new research regarding the calculation of statistical power and sizing of analysis samples
- Provided more guidance about estimating impacts of behavior-based programs on participation in other energy efficiency programs
- Edited the text in various places to improve organization or to clarify concepts and recommendations.

Table of Contents

List of Figures

List of Tables

1 Measure Description

Residential behavior-based (BB) programs use strategies grounded in the behavioral and social sciences to influence household energy use. These may include providing households with real-time or delayed feedback about their energy use; supplying energy efficiency education and tips; rewarding households for reducing their energy use; comparing households to their peers; and establishing games, tournaments, and competitions.[1] BB programs often target multiple energy end uses and encourage energy savings, demand savings, or both. Savings from BB programs are usually a small percentage of energy use, typically less than 5%.[2]

Utilities introduced the first large-scale residential BB programs in 2008. Since then, dozens of utilities have offered these programs to their customers.[3] Although program designs differ, many share these features:

- They are implemented as randomized experiments wherein eligible homes are randomly assigned to treatment or control groups.
- They are large scale by energy efficiency program standards, targeting thousands of utility customers.
- They provide customers with analyses of their historical consumption, energy savings tips, and energy efficiency comparisons to neighboring homes, either in personalized home reports or through a web portal, or offer incentives for savings energy.
- They are typically implemented by outside vendors.[4]

Utilities will continue to implement residential BB programs as large-scale, randomized control trials (RCTs); however, some are now experimenting with alternative program designs that are smaller scale; involve new communication channels such as the web, social media, and text messaging; or that employ novel strategies for encouraging behavior change (for example, Facebook competitions).[5] These programs will create new evaluation challenges and may require different evaluation methods than those currently employed to verify any savings they generate. Quasi-experimental methods, however, require stronger assumptions to yield valid savings estimates and may not measure savings with the same degree of validity and accuracy as randomized experiments.

[1] See Ignelzi et al. (2013) for a classification and descriptions of different BB intervention strategies and Mazur-Stommen and Farley (2013) for a survey and classification of current BB programs. Also, a Minnesota Department of Commerce, Division of Energy Resources white paper (2015) defines, classifies, and benchmarks behavioral intervention strategies.
[2] See Allcott (2011), Davis (2011), and Rosenberg et al. (2013) for savings estimates from residential BB programs.
[3] See the 2013 Consortium for Energy Efficiency (CEE) database for a list of utility behavior programs; it is available for download: http://library.cee1.org/content/2013-behavior-program-summary-public-version.
[4] Vendors that offer residential BB programs include Aclara, C3 Energy, ICF, Oracle Utilities (Opower), Simple Energy, and Tendril.
[5] The 2013 CEE database includes descriptions of many residential BB programs with alternative designs such community-focused programs, college dormitory programs, K-12 school programs, and programs relying on social media.

2 Application Conditions of Protocol

This protocol recommends the use of RCTs or randomized encouragement designs (REDs) for estimating savings from BB programs. A significant body of research indicates that randomized experiments result in unbiased and robust estimates of program energy and demand savings. Moreover, recently evaluators have conducted studies comparing the accuracy of savings estimates from randomized experiments and quasi-experiments or observational studies. These comparisons suggest that randomized experiments produce the most accurate savings estimates.[6]

This protocol applies to BB programs that satisfy the following conditions:[7]

- Residential utility customers are the target.
- Energy or demand savings are the objective.
- An appropriately sized analysis sample can be constructed.
- Treated customers can be identified and accurate energy use measurements for sampled units are available.
- It must be possible to isolate the treatment effect when measuring savings.

This protocol applies only to residential BB programs. Although the number of nonresidential BB programs is growing, utilities offer a larger number of residential BB programs and to a much larger number of residential customers.[8] As evaluators accumulate more experience, the National Renewable Energy Laboratory (NREL) could expand this protocol to cover nonresidential programs to which similar evaluation methods are applicable.

This protocol also addresses best practices for estimating energy and demand savings. There are no significant conceptual differences between measuring energy savings and measuring demand savings when interval data are available; thus, evaluators can apply the algorithms in this protocol for calculating BB program savings to either. The protocol does not directly address the evaluation of other BB program objectives, such as increasing utility customer satisfaction, educating customers about their energy use, or increasing awareness of energy efficiency.[9] But

---

[6] Allcott (2011) compares RCT difference-in-differences (DiD) savings estimates with quasi-experimental simple differences and DiD savings estimates for several home energy reports programs. He found large differences between the RCT and quasi-experimental estimates. Also, Baylis et al. (2016) analyzed data from a California utility time-of-use and critical peak pricing pilot program and found that RCT produced more accurate savings estimates than quasi-experimental methods such as DiD and propensity score matching that relied on partly random but uncontrolled variation in participation.

[7] As discussed in the "Considering Resource Constraints" section of the UMP *Chapter 1: Introduction*, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

[8] Evaluators may be able to apply the methods recommended in this protocol to the evaluation of some nonresidential BB programs. For example, Pacific Gas and Electric (PG&E) offers a Business Energy Reports Program, which it implemented as an RCT (Seelig 2013). Also, Xcel Energy implemented a business energy reports program as an RCT (Stewart 2013b). Other nonresidential BB programs may not lend themselves to evaluation by randomized experiment. For example, many strategic energy management programs enroll large industrial customers with unique production and energy consumption characteristics for which a randomized experiment would not be feasible (NREL 2017).

[9] Process evaluation objectives may be important, and omission of them from this protocol should not be interpreted as a statement that these objectives should not be considered by program administrators.

these program outcomes could be studied in a complementary fashion alongside the energy savings.

This protocol also requires that the analysis sample be large enough to detect the expected savings with a high degree of confidence. Because most BB programs result in small percentage savings, a large sample size is required to detect savings. This protocol does not address evaluations of BB programs with a small number of participants.

Finally, this protocol requires that the energy use of participants or households affected by the program (for the treatment and control groups) can be clearly identified and measured. Typically, the analysis unit is the household; in this case, treatment group households must be identifiable and individual household energy use must be measurable. However, depending on the BB program, the analysis units may not be households. For example, for a BB program that generates an energy competition between hundreds of housing floors at a university, the analysis unit may be floors; in this case, the energy use measurement of individual floors must be available.

The characteristics of BB programs that *do not* determine the applicability of the evaluation protocol include:

- Whether the program is opt-in or opt-out[10]
- The specific behavior-modification theory or strategy
- The channel(s) through which program information is communicated.

Although this protocol strongly recommends RCTs or REDs, it also recognizes that implementing these methods may not always be feasible. Government regulations or program designs may prevent the utilization of randomized experiments for evaluating BB programs. In these cases, evaluators must employ quasi-experimental methods, which require stronger assumptions than do randomized experiments to yield valid savings estimates.[11] If these assumptions are violated, quasi-experimental methods may produce biased results. The extent of the biases in the estimates is not knowable *ex ante*, so results will be less reliable. Because there is currently not enough evidence of quasi-experimental methods that perform well, this protocol refrains from recommending non-RCT evaluation methods. As noted above, studies have found quasi-experiments produce less accurate savings estimates than randomized experiments. A good reference for applying quasi-experimental methods to BB program evaluation is State and Local Energy Efficiency Action (SEE Action) (2012) or Cappers et al. (2013). As more evidence accumulates about the efficacy of quasi-experiments, the National Renewable Energy Laboratory may update this protocol as appropriate.

2.1 Examples of Protocol Applicability

Examples of residential BB programs for which the evaluation protocol applies follow:

[10] In opt-in programs, customers enroll or select to participate. In opt-out programs, the utility enrolls the customers, and the customers remain in the program until they opt out. An example opt-in program is having a utility web portal with home energy use information and energy efficiency tips that residential customers can use if they choose. An example opt-out program is sending energy reports to utility selected customers.

[11] For example, Harding and Hsiaw (2012) use variation in timing of adoption of an online goal-setting tool to estimate savings from the tool.

- **Example 1:** A utility sends energy reports encouraging conservation steps to thousands of randomly selected residential customers.
- **Example 2:** Several hundred residential customers enroll in a Wi-Fi-enabled thermostat pilot program offered by the utility.
- **Example 3:** A utility invites thousands of residential customers to use its web portal to track their energy use in real time, set goals for energy saving, find ideas about how to reduce their energy use, and receive points or rewards for saving energy.
- **Example 4**: A utility sends voice, text, and email messages to thousands of residential utility customers encouraging—and providing tips for— reducing energy use during an impending peak demand event.

Examples of programs for which the protocol does not apply follow:

- **Example 5:** A utility uses a mass-media advertising campaign that relies on radio and other broadcast media to encourage residential customers to conserve energy.
- **Example 6:** A utility initiates a social media campaign (for example, using Facebook or Twitter) to encourage energy conservation.
- **Example 7:** A utility runs a pilot program to test the savings from in-home energy-use displays, and enrolls too few customers to detect the expected savings.
- **Example 8:** A utility runs a BB program in a large college dormitory to change student attitudes about energy use. The utility randomly assigns some rooms to the treatment group. The dorm is master-metered.

The protocol does not apply to Example 5 or Example 6 because the evaluator cannot identify who received the messages. The protocol does not apply to Example 7 because too few customers are in the pilot to accurately detect energy savings. The protocol does not apply to Example 8 because energy-use data are not available for the specific rooms in the treatment and control groups.

3 Savings Concepts

The protocol recommends RCTs and REDs to develop unbiased and robust estimates of energy or demand savings from BB programs that satisfy the applicability conditions described in Section 2. Unless otherwise noted, all references in this protocol to savings are to net savings.

Section 3.1 defines some key concepts and Section 3.2 describes specific evaluation methods.

3.1 Definitions

The following key concepts are used throughout this protocol.

Control group. In an experiment, the control group comprises subjects (for example, utility customers) who do not receive the program intervention or treatment.

Experimental design.[12] Randomized experimental designs rely on observing the energy use of subjects who were randomly assigned to program treatments or interventions in a controlled process.

External validity. Savings estimates are externally valid if evaluators can apply them to different populations or different time periods from those studied.

Internal validity. Savings estimates are internally valid if the savings estimator is expected to equal the causal effect of the program on consumption.

Opt-in program. Utilities use opt-in BB programs if the customers must agree to participate, and the utility cannot administer treatment without consent.

Opt-out program. Utilities use opt-out BB programs if customers need not agree to participate. The utility can administer treatment without consent, and customers remain enrolled until they ask the utility to stop the treatment.

Quasi-experimental design. Quasi-experimental designs rely on a comparison group who is not obtained via random assignment. Such designs observe energy use and determine program treatments or interventions based on factors that may be partly random but not controlled.

Randomized Control Trial (RCT). An RCT uses random variation in which subjects are exposed to the program treatment to obtain an estimate of the treatment effect. By randomly assigning subjects to treatment, an RCT controls for factors that could confound measurement of the treatment effect. An RCT is expected to yield an unbiased estimate of program savings. Evaluators randomly assign subjects from a study population to a treatment group or a control group. Subjects in a treatment group receive one program treatment (there may be multiple treatments), while subjects in the control group receive no treatment. The RCT ensures that receiving the treatment is uncorrelated with the subjects' pre-treatment energy use, and that evaluators can attribute any difference in energy use between the groups to the treatment.

[12] When this protocol uses the term randomized experiments, it refers to RCTs or REDs, not other experimental evaluation approaches such as natural experiments or quasi-experiments.

Randomized Encouragement Design (RED). In an RED, evaluators randomly assign subjects to a treatment group that receives *encouragement* to participate in a program or to a control group who does not receive encouragement. The RED yields an unbiased estimate of the effect on energy use of encouraging energy-efficient behaviors and the effect on customers who participate because of the encouragement.

Treatment. A treatment is an intervention administered through the BB program to subjects in the treatment group. Depending on the research design, the treatment may be a program intervention or encouragement to accept an intervention.

Treatment effect. This is the effect of the BB program intervention(s) on energy use for a specific population and time period.

Treatment group. The treatment group includes subjects who receive the treatment.

3.2 Randomized Experimental Research Designs

This section outlines the application of randomized experiments for evaluating BB programs. The most important benefit of an RCT or RED is that, if carried out correctly, the experiment results in an unbiased estimate of the program's causal impact.[13] Unbiased savings estimates have internal validity. A result is internally valid if the evaluator can expect the value of the estimator to equal the savings caused by the program intervention. The principal threat to internal validity in BB program evaluation derives from potential selection bias about who receives a program intervention. RCTs and REDs yield unbiased savings estimates because they ensure that receiving the program intervention is uncorrelated with the subjects' energy use.

Randomized experiments may yield savings estimates that are applicable to other populations or time periods, making them externally valid. Whether savings have external validity will depend on the specific research design, the study population, and other program features.[14] Program administrators should exercise caution in applying BB program savings estimates for one population to another or to the same population at a later time, since differences in population characteristics, weather, or naturally-occurring efficiency can cause savings to change.

A benefit of field experiments is their versatility: evaluators can apply them to a wide range of BB programs regardless of whether they are opt-in or opt-out programs. Evaluators can apply randomized experiments to any program where the objective is to achieve energy or demand savings; evaluators can construct an appropriately sized analysis sample; and accurate measurements of the energy use of sampled units are available.

Randomized experiments generally yield highly robust savings estimates that are not model dependent; that is, they do not depend on the specification of the model used for estimation.

The choice of whether to use an RCT or RED to evaluate program savings should depend on several factors, including whether it is an opt-in or opt-out program, the expected number of

[13] List (2011) describes many of the benefits of employing randomized field experiments.
[14] Allcott (2015) analyzes the external validity of savings estimates from evaluations of 111 RCTs of home energy reports programs in the United States and shows that the first utilities implementing the programs achieved higher savings than utilities that implemented such programs subsequently.

program participants, and the utility's tolerance for subjecting customers to the requirements of an experiment. For example, using an RCT for an opt-in program might require delaying or denying participation for some customers. A utility may prefer to use an RED to accommodate all the customers who want to participate.

Implementing an RCT or RED design requires upfront planning. Program evaluation must be an integral part of the program planning process, which is evident in the randomized experiment research design descriptions in Section 3.3.

3.3 Basic Features

This section outlines several types of RCT research designs, which are simple but extremely powerful research tools. The core feature of RCT is the random assignment of study subjects (for example, utility customers, floors of a college dormitory) to a treatment group that receives or experiences an intervention or to a control group that does not receive the intervention.

Section 3.3.1 outlines some common features of RCTs and discusses specific cases.

3.3.1 Common Features of Randomized Control Trial Designs

The key requirements of an RCT are incorporated into the following steps:

1. **Identify the study population:** The program administrator screens the utility population if the program intervention is offered to certain customer segments only, such as single-family homes. Programs designers can base eligibility on dwelling type (for example, single family, multifamily), geographic location, completeness of recent billing history, heating fuel type, utility rate class, or other energy use characteristics.
2. **Determine sample sizes:** The numbers of subjects to assign to the treatment and control groups depend on the type of randomized experiment (for example, REDs and opt-out RCTs generally require more customers) and hypothesized savings. The number of subjects assigned to the treatment versus control groups should be large enough to detect the hypothesized program effect with sufficient probability, though it is not necessary for the treatment and control groups to be equally sized.[15]

   Evaluators can use a statistical power analysis to determine the number of subjects required. This results in minimum sample sizes for the treatment and control groups as a function of the hypothesized program effect, the coefficient of variation of energy use, the specific analysis approach that will be used (for example, simple differences of means, a repeated measure analysis where there are multiple observations of energy consumption at different time periods for the same subject [aka, panel analysis]), and tolerances for Type I and Type II statistical errors.[16] Most statistical software (including SAS, STATA, and R) now include packages for performing statistical power analyses. It

[15] The number of subjects in the treatment group may also depend on the savings goal for the program.

[16] A Type I error occurs when a researcher rejects a null hypothesis that is true. Statistical confidence equals 1 minus the probability of a Type I error. A Type II error occurs when a researcher accepts a null hypothesis that is false. Many researchers agree that the probability of a 5% Type I error and a 20% Type II error is acceptable. See List et al. (2010).

is not uncommon for BB programs with expected savings of less than 3% to require thousands of subjects in the treatment and control groups.[17]

An important component of the random assignment process is to verify that the treatment and control groups are statistically equivalent or balanced in their observed covariates. At a minimum, evaluators should check before the intervention for statistically significant differences in average pre-treatment energy use and in the distribution of pre-treatment energy use between treatment and control homes.

3. **Randomly assign subjects to treatments and control:** Study subjects should be randomly assigned to treatment and control groups. To maximize the credibility and acceptance of BB program evaluations, this protocol recommends that a qualified independent third party perform the random assignment. Also, to preserve the integrity of the experiment, customers must not choose their assignments. The procedure for randomly assigning subjects to treatment and control groups should be transparent and well documented.
4. **Administer the treatment:** The intervention must be administered to the treatment group and withheld from the control group. To avoid a Hawthorne effect, in which subjects change their energy use in response to observation, control group subjects should receive minimal information about the study. Depending on the research subject and intervention type, the utility may administer treatment once or repeatedly and for different durations. However, the treatment period should be long enough for evaluators to observe any effects of the intervention.
5. **Collect data:** Data must be collected from all study subjects, not only from those who chose to participate or only from those who participated for the whole study or experiment.

   Preferably, evaluators collect multiple pre- and post-treatment energy use measurements. Such data enable the evaluator to control for time-invariant differences in average energy use between the treatment and control groups to obtain more precise savings estimates. Step 6 discusses this in further detail.
6. **Estimate savings:**[18] Evaluators should calculate savings as the difference in energy use or difference-in-differences (DiD) of energy use between the subjects who were initially assigned to the treatment versus the control group. To be able to calculate an unbiased savings estimate, evaluators must compare the energy use from the entire group of subjects who were originally randomly assigned to the treatment group to the entire group of subjects who were originally randomly assigned to the control group. For example, the savings estimate would be biased if evaluators used only data from utility customers in the treatment group who chose to participate in the study.

   The difference in energy use between the treatment and control groups, usually called an intent-to-treat (ITT) effect, is an unbiased estimate of savings because subjects were

[17] EPRI (2010) illustrates that, all else equal, repeated measure designs, which exploit multiple observations of energy use per subject both before and after program intervention, require smaller analysis sample sizes than other types of designs.
[18] This protocol focuses on estimating average treatment effects; however, treatment effects of behavior programs may be heterogeneous. Costa and Kahn (2010) discuss how treatment effects can depend on political ideology and Allcott (2011) discusses how treatment effects can depend on pretreatment energy use.

randomly assigned to the treatment and control groups. The effect is an ITT because, in contrast to many randomized clinical medical trials, ensuring that treatment group subjects in most BB programs comply with the treatment is impossible. For example, some households may opt out of an energy reports program, or they may fail to notice or simply ignore the energy reports. Thus, the effect is ITT, and the evaluator should base the results on the initial assignment of subjects to the treatment group, whether or not subjects actually complied with the treatment.

The savings estimation approach should be well documented, transparent, and performed by an independent third party.

3.4 Common Designs

This section describes some of the RCT designs commonly used in BB programs.

3.4.1 Randomized Control Trial With Opt-Out Program Design

One common type of RCT includes the option for treated subjects to opt out of receiving the program treatment. This design reflects the most realistic description of how most BB programs work. For example, in energy reports programs, some treated customers may ask the utility to stop sending them reports.

Figure 1 depicts the process flow of an RCT in which treated customers can opt out of the program. In this illustration, the utility initially screened utility customers to refine the study population.[19]

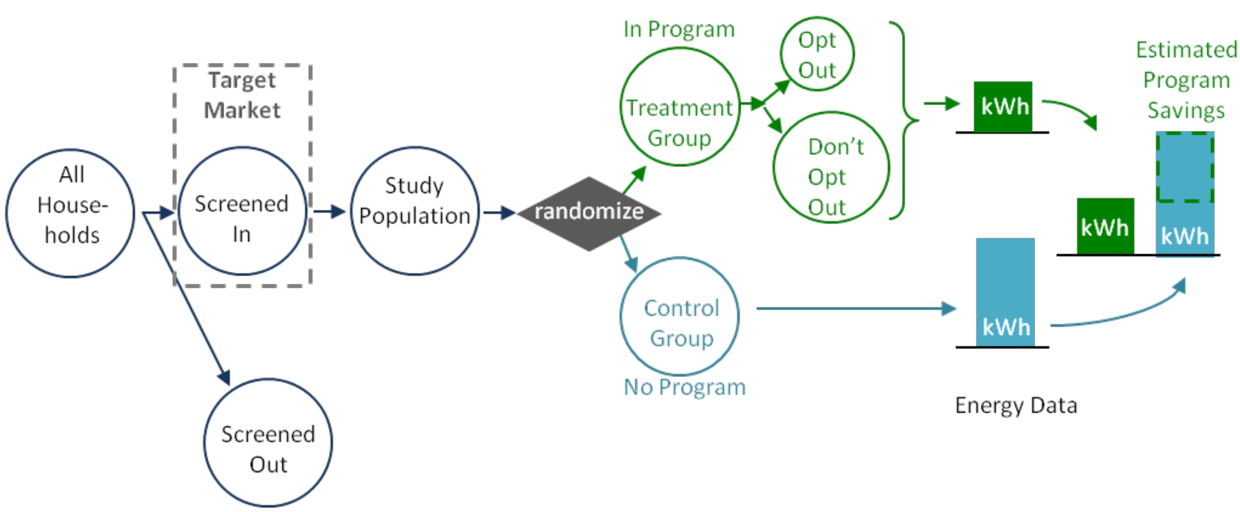


Figure 1. Illustration of RCT with opt-out program design

Customers who pass the screening constitute the study population or sample frame. The savings estimate will apply to this population. Alternatively, the utility may want to study only a sample of the screened population, in which case a third party should sample randomly from the study population. The analysis sample must be large enough to meet the minimum size requirement for

[19] This graphic and the following ones are variations of those that appeared in SEE Action (2012). A coauthor of the SEE Action report and the creator of that reports' figures is one of the authors of this protocol.

the treatment and control groups. The program savings goals and desired statistical power will determine the size of the treatment group.

The next steps in an RCT with opt-out program design are to (1) randomly assign subjects in the study population to the program treatment and control groups, (2) administer the program treatments, and (3) collect energy use data.

The distinguishing feature of this randomized experimental design is that customers can opt out of the program. As Figure 1 shows, evaluators should include opt-out subjects in the energy savings analysis to ensure unbiased savings estimates. Evaluators can then calculate savings as the difference in average energy use between treatment group customers, including opt-out subjects and control group customers. Removing opt-out subjects from the analysis would bias the savings estimate because identifying subjects in the control group who would have also opted out had they received the treatment is impossible. The resulting savings estimate is therefore an average of the savings of treated customers who remain in the program and of customers who opted out.

Depending on the type of BB program, the percentage of customers who opt out may be small, and may not affect the savings estimates significantly (for example, few customers generally opt out of energy reports programs).

3.4.2 Randomized Control Trial With Opt-In Program Design

Utilities must have consent from customers to administer some program interventions. Examples include web-based home audit or energy consumption tools; programmable, communicating thermostats with wireless capability; online class about energy rates and efficiency; or in-home displays. All these interventions require that customers opt in to the program. These interventions contrast with interventions such as home energy reports that can be administered to subjects without their agreement.

An opt-in RCT (Figure 2) can accommodate the necessity for customers to opt in to some BB programs. This design results in an unbiased estimate of the ITT effect for customers who opt in to the program. The estimate of savings will have internal validity; however, it will not have external validity because it will not apply to subjects who do not opt in.

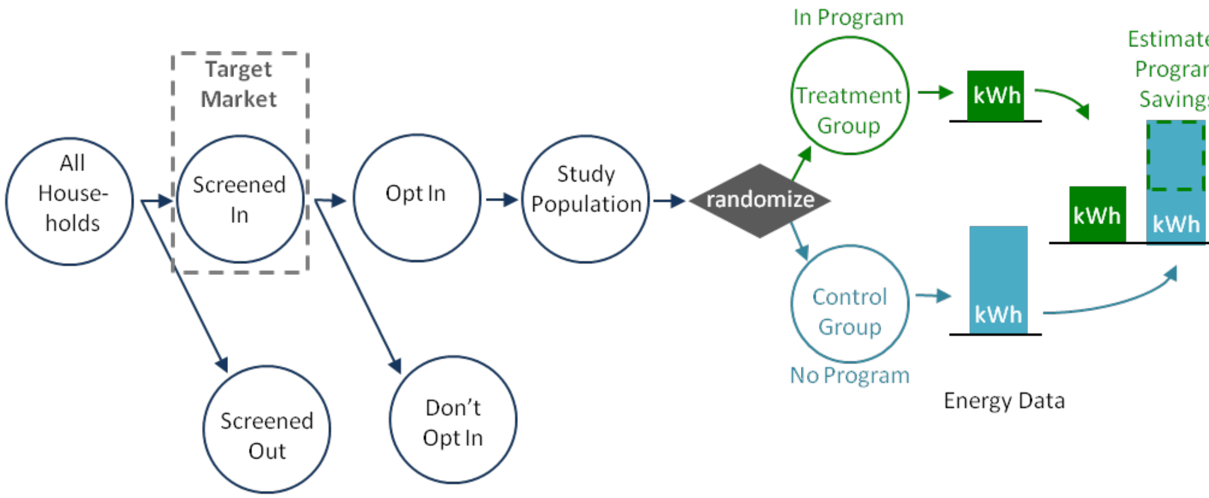


Figure 2. Illustration of RCT with opt-in program design

Implementing opt-in RCTs is very similar to implementing opt-out RCTs. The first step, screening utility customers for eligibility to determine the study population, is the same. The next step is to market the program to eligible customers. Some eligible customers may then agree to participate. Then, an independent third party randomly assigns these customers to either a treatment group that receives the intervention or a control group that does not. The utility delays or denies participation in the program to customers assigned to the control group. Thus, only customers who opted in and were assigned to the treatment group will receive the treatment.

Randomizing only opt-in customers ensures that the treatment and control groups are equivalent in their energy use characteristics. In contrast, other quasi-experimental approaches, such as matching participants to nonparticipants, cannot guarantee either this equivalence or the internal validity of the savings estimates.

After the random assignment, the opt-in RCT proceeds the same as an RCT with opt-out subjects: the utility administers the intervention to the treatment group. The evaluator collects energy use data from the treatment and control groups, then estimates energy savings as the difference in energy use between the groups. The evaluator does not collect energy use data for customers who do not opt in to the program.

An important difference between the opt-in RCTs and RCTs with opt-out subjects is how to interpret the savings estimates. In the RCT with opt-out subjects, the evaluator bases the savings estimate on a comparison of the energy use between treatment and control groups, which pertains to the entire study population. In contrast, in the opt-in RCT, the savings estimate pertains to the subset of customers who opted into the program, and the difference in energy use represents the treatment effect on customers who opted in to the program. Opt-in RCT savings estimates have internal validity; however, they do not apply to customers who did not opt in to the program.

3.4.3 Randomized Encouragement Design

For some opt-in BB programs, delaying or denying participation to some customers may be undesirable. In this case, neither the opt-out nor the opt-in RCT design would be appropriate, and this protocol recommends an RED. Instead of randomly assigning subjects to receive or not receive an intervention, a third party randomly assigns them to a treatment group that is *encouraged* to accept the intervention (that is, to participate in a program or adopt a measure), or to a control group that does not receive encouragement. Examples of common kinds of encouragement include direct paper mail or e-mail informing customers about the opportunity to participate in a BB program. Customers who receive the encouragement can refuse to participate, and, depending on the program design, control group customers who learn about the program may be able to participate.

The RED yields an unbiased estimate of the effect of encouragement on energy use and, depending on the program design, can also provide an unbiased estimate of either the effect of the intervention on customers who accept it because of the encouragement or the effect of the intervention on all customers who accept it. A necessary condition for an RED to produce an unbiased estimate of savings from the BB intervention is that the encouragement only affects energy consumption for those customers that take up the BB intervention, and it does not affect the energy consumption for customers who receive the encouragement but do not take up the BB intervention. For example, the RED must be such that customers who receive a direct mailing encouraging them to log into a website with personalized energy efficiency recommendations only save energy if they decide to log into the site; the mailing itself must not cause the customer to save energy if the customer never logs on. If the encouragement causes customers to save energy, it may be impossible to isolate the savings from the intervention. Programs designed as an RED should try to design and distribute encouragement materials that do not affect consumption. If evaluators expect that the encouragement will cause energy savings, they can send the same or similar messaging but without a program enrollment option to the control group or to a second randomized control group. Evaluators could use the second randomized control group to test whether the encouragement produces savings and to estimate the savings from the encouragement.

Figure 3 illustrates the process flow for a program using an RED. As with the RCT with opt-out and opt-in RCT, the first two steps are to identify the sample frame and select a study population. Next, like the RCT with opt out, a third party randomly assigns subjects to a treatment group, which receives encouragement, or to a control group, which does not. For example, a utility might employ a direct mail campaign that encourages treatment group customers to use an online audit tool. The utility would administer the intervention to treatment group customers who opt-in. Although customers in the control group did not receive encouragement, some may learn about the program and decide to sign up. The program design shown in Figure 3 allows for control group customers to receive the behavioral intervention.

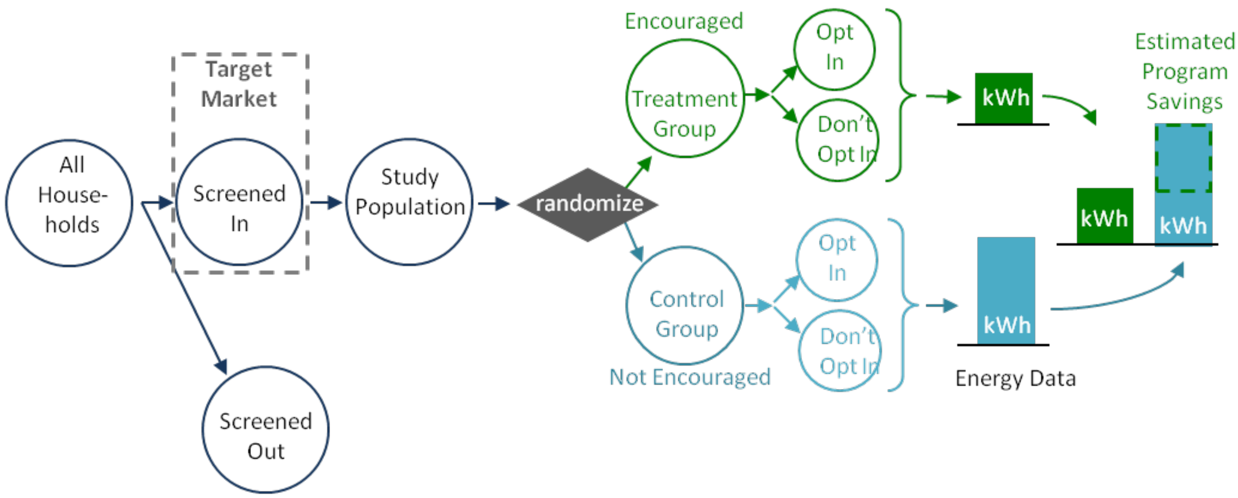


Figure 3. Illustration of RED program design

In Figure 3, the difference in energy use between homes in the treatment and control groups is an estimate of savings from the encouragement, not from the intervention. However, evaluators can also use the difference in energy use to estimate savings for customers who accept the intervention because of the encouragement. To see this, consider that the study population comprises three types of subjects: (1) always takers, or those who would accept the intervention whether encouraged or not; (2) never takers, or those who would never accept the intervention even if encouraged; and (3) compliers, or those who would accept the intervention only if encouraged. Compliers participate only after receiving the encouragement.

Because eligible subjects are randomly assigned to groups depending on whether they receive encouragement, the treatment and control groups are expected to have equal frequencies of always takers, never takers, and compliers. After treatment, the only difference between the treatment and control groups is that compliers in the treatment group accept the treatment and compliers in the control group do not. In both groups, always takers accept the treatment and never takers always refuse the treatment. Therefore, the difference in energy use between the groups reflects the treatment effect of encouragement on compliers (known as the local average treatment effect [LATE]).

Furthermore, for the study to have enough statistical power to detect the expected effect, there must be very large encouraged and non-encouraged groups relative to an RCT or quasi-experimental design and/or a high proportion of compliers in the treatment group; a power calculation should be done to ensure that there are enough customers in the encouraged and non-encouraged groups to produce significant savings estimates for the expected take-up rate.[20]

To estimate the effect of the intervention on compliers, evaluators can either employ instrumental variables (IV), using the random assignment of customers to receive encouragement as an instrument for the customer's decision to accept the intervention (that is, participate). The IV approach is presented in Section 4.3. Another option is that evaluators can scale the treatment effect of the encouragement by the difference between treatment and control groups in the

[20] For an example of a power calculation for REDs, see Fowlie (2010).

percentage of customers who receive the intervention (note that in this equation, if the non-encouraged customers are not allowed to take up the treatment, the second term in the denominator will be zero):[21]

1/(% of encouraged customers who accepted –
% of non-encouraged customers who accepted)

If customers in the control group are permitted to participate if they find out about the treatment even though they did not receive encouragement, the LATE does not capture the program effect on always takers. (Note, however, in most programs, the control group is not permitted to take up the treatment). If customers in the control group are permitted to participate, the LATE may differ from the average treatment effect unless the savings from the intervention is the same for compliers and always takers. However, the LATE will be equal to the average treatment effect if the control group customers (non-encouraged customers) are not permitted to take up the treatment.

For BB programs with REDs that do not permit control group customers to participate, evaluators can estimate the treatment effect on the treated (TOT). The TOT is the effect of the program intervention on all customers who accept the intervention. In this case, the difference in energy use between the treatment and control groups reflects the impact of the encouragement on the always takers and compliers in the treatment group. Scaling the difference by the inverse of the percentage of customers who accepted the intervention yields an estimate of the TOT impact.[22]

Successful application of an RED requires that compliers constitute a percentage of the encouraged population that is sufficiently large given the number of encouraged customers.[23] If the RED generates too few compliers, the effects of the encouragement and receiving the intervention cannot be precisely estimated. Therefore, before employing an RED, evaluators should ensure that the sample size is sufficiently large and that the encouragement will result in the required number of compliers. If the risk of an RED generating too few compliers is significant, evaluators may want to consider alternative approaches, including quasi-experimental methods.

3.4.4 Persistence Design

Studies of home energy reports programs show that program savings persist while homes continue to receive reports. However, utilities and regulators may want to know what happens to BB program savings after the behavioral intervention ends. They may wish to measure whether their savings persist after the utility stops sending reports and for how long, as well as the rate of the savings "decay." As Allcott and Rodgers (2014) demonstrate, the rate of savings decay after treatment ends has significant implications for the performance of efficiency program portfolios

[21] This approach of estimating savings from the intervention because of encouragement assumes zero savings for customers who received encouragement but did not accept the intervention. If encouraged customers who did not accept the intervention reduced their energy use in response to the encouragement, the savings estimate for compliers will be biased upward.

[22] If the effect of program participation is the same for compliers as for others, those who would have participated without encouragement (always takers) and those who do not participate (never takers), the RED will yield an unbiased estimate of the population average treatment effect.

[23] For an example of the successful application of an RED, see SMUD (2013).

and measuring cost effectiveness of BB programs. Initial studies of home energy reports programs indicate that some portion of savings may persist after the treatment stops, although further research is needed.[24]

This protocol recommends that evaluators employ RCTs to estimate the persistence of BB program savings after participants stop receiving the intervention. The application of an RCT to a savings persistence study proceeds similarly to the application of RCTs previously discussed.

The utility is assumed to implement the BB program as an RCT with opt-out design; that is, customers from the study population were randomly assigned to a treatment group that received an intervention or to a control group that did not. Customers are able to opt out of the program (see Figure 1).

The persistence study starts with identifying the study population, in this case, the population of treated customers who received the intervention. The utility may choose to screen this population and study persistence by energy use or by socio-demographic characteristics. The persistence study population must include customers who opted out, because evaluators will need to make energy use comparisons between the persistence study population and the original control group, which includes customers who would have opted out.

The next step is to randomly assign customers in the persistence study population to one of two groups. Customers in the "discontinued treatment" group will stop receiving the intervention; customers in the "continued treatment" group will continue receiving the intervention. The utility then administers the study and collects energy consumption data after sufficient time has passed to observe the persistence effects.

To estimate savings after the end of treatment, the evaluator compares the energy consumption of customers in the discontinued treatment group with the energy consumption of customers in the original control group. The difference represents the post-treatment savings for customers who no longer received the intervention.

To estimate savings persistence, the evaluator compares the savings of the continued and discontinued treatment groups after the end of treatment. The ratio of the discontinued group savings to the continued group savings is the percentage of savings that persists after treatment ends. Savings decay is the difference in savings between the continued and discontinued treatment groups, and the savings decay rate is the average savings decay per period.

3.5 Evaluation Benefits and Implementation Requirements of Randomized Experiments

This protocol strongly recommends the use of randomized field experiments (RCTs or REDs) for evaluating residential BB programs. Table 1 summarizes the benefits and requirements of evaluating BB programs using RCTs and REDs, as described in Sections 3.1–3.4.

[24] Studies show that savings may persist after treatment stops (Allcott and Rodgers 2014; Brattle 2012; SMUD 2011; PSE 2012; Khawaja and Stewart 2014; Olig and Young 2016; and Skumatz 2016). Allcott and Rodgers (2014) estimate a savings decay rate of about 19% per year. Brandon et al. (2017) provide evidence that up to half of Home Energy Report savings persistence is attributable to physical capital improvements to homes.

Table 1. Benefits and Implementation Requirements of Randomized Experiments

Evaluation Benefits	Implementation Requirements
• Yield unbiased, valid estimates of causal program impacts, resulting in a high degree of confidence in the savings • Yield savings estimates that are robust to changes in model specification • Are versatile, and can be applied to opt-out and opt-in BB programs • Are widely accepted as the "gold standard" of good program evaluations • Result in transparent analysis and evaluation • Can be designed to test specific research questions such as persistence of savings after treatment ends	• An appropriately sized analysis sample • Accurate energy use measurements for sampled units • Advance planning and early evaluator involvement in program design • Restricted participation or program marketing to randomly selected customers

The principal benefit of randomized experiments is that they yield unbiased and robust estimates of program savings. They are also versatile, widely accepted, and straightforward to analyze. The principal requirements for implementing randomized experiments include the availability of accurate energy use measurements and a sufficiently large analysis study population.[25]

Also, this protocol specifically recommends REDs or RCTs for estimating BB program savings as both designs yield unbiased savings estimates. The choice of RED or RCT will depend primarily on program design and implementation considerations, in particular, whether the program has an opt-in or opt-out design. RCTs work well with opt-out programs such as residential energy reports programs. Customers who do not want to receive reports can opt out at any time without adversely affecting the evaluation. RCTs also work well with opt-in programs for which customer participation can be delayed (for example, customers are put on a "waiting list") or denied. For situations in which delaying or denying a certain subset of customers is impossible or costly, REDs may be more appropriate. REDs can accommodate all interested customers, but have the disadvantages of requiring larger analysis samples, two analysis steps to yield a direct estimate of the behavioral intervention's effect on energy use, and a high proportion of compliers among encouraged customers.

Table 2 lists some issues to consider when choosing an RCT or RED.

[25] A frequent objection to the use of randomized experiments is that some utility customers may not have the opportunity to participate in a program. However, programs are often limited to a certain subset of customers; for example, a program may start out as limited to customers in a certain county or other geographic location. REDs allow any customers who would like to participate the opportunity to do so, even if they are in the control group. In our view, limiting the availability of the program to certain customers in RCTs is done with the worthy objective of advancing the utility's knowledge of program savings effects and making future allocation of scarce efficiency resources more optimal.

Table 2. Considerations in Selecting a Randomized Experimental Design

Experimental Design	Evaluation Benefits	Implementation and Evaluation Requirements
RCT	• Yields unbiased, robust, and valid estimates of causal program impacts, resulting in a high degree of confidence in the savings • Simple to understand • Works well with opt-out programs • Works well with opt-in programs if customers can be delayed or denied	• May require delaying or denying participation of some customers if program requires customers to opt in
RED	• Yields unbiased, robust, and valid estimates of causal program impacts, resulting in a high degree of confidence in the savings • Can accommodate all customers interested in participating • Works well with opt-in and opt-out programs	• More complex design and harder to understand • Requires a more complex analysis • Requires larger analysis sample • Requires a proportion of compliers that is sufficient given the number of encouraged customers to estimate savings • Encouragement to participate should not cause customers to save energy

4 Savings Estimation

Evaluators should estimate BB program savings as the difference in energy use between treatment and control group subjects in the analysis sample. Energy savings for a household in the BB program is the difference between the energy the household used and the energy the household would have used if it had not participated. However, the energy use of a household cannot be observed under two different states. Instead, to estimate savings, evaluators should compare the energy use of households in the treatment group to that of a group of households that are statistically the same but did not receive the treatment (the homes randomly assigned to the control group). In a randomized experiment, assignment to the treatment is random; thus, evaluators can expect control group subjects to use the same amount of energy that the treatment group would have used without the treatment. The difference in their energy use will therefore be an unbiased estimate of energy savings.

Savings can be estimated using energy use data from the treatment period only or from before and during the treatment. If energy use data from only the treatment period are used, evaluators estimate the savings as a simple difference (D). If the analysis also controls for energy use before the treatment, evaluators can estimate the savings as a DiD or as a simple difference that controls for pre-treatment energy consumption. The approach that estimates savings conditional on pre-treatment consumption is sometimes referred to as a "post-only model."[26] The availability of energy use data for the period before the treatment will determine the approach, but incorporating pre-treatment consumption data in the analysis is strongly advised when such data are available.

Both approaches result in unbiased estimates of savings (that is, in expectation, the two methods are expected to yield an estimate equal to the true savings). However, estimators using pre-treatment data generally result in more precise savings estimates (that is, the estimators using pre-treatment data will have a smaller standard error) as it accounts for time-invariant energy use that contribute significantly to the variance of energy use between subjects.[27]

Evaluators should collect at least one full year of historical energy use data (the 12 months immediately before the program start date) to ensure baseline data fully reflect seasonal energy use effects.

Regulators usually determine the frequency of program evaluation. Although requirements vary between jurisdictions, most BB programs are evaluated once per year. Annual evaluation will likely be necessary for the first several years of many BB programs such as home energy reports programs because savings tend to increase for several years before leveling off. However, some

---

[26] The model with pre-treatment consumption control variables is a significantly more efficient estimator (that is, it is expected to have smaller variance) than the DiD estimator when the model errors are independent and identically distributed or when serial correlation of consumption is low (Burlig, Preonas, and Woerman 2017). This model is more efficient because it uses one degree of freedom rather than multiple degrees of freedom—one for each study subject—to account for between-subject differences in consumption. However, when serial correlation of customer consumption is high, there is little or no gain in efficiency over the fixed effects the DiD approach.

[27] Post-only or DiD estimation with customer fixed effects also accounts for differences in mean energy use between treatment and control group subjects that are introduced when subjects are randomly assigned to the treatment or control group. Evaluators may not expect such differences with random assignment; however, these differences may nevertheless arise.

program administrators may desire measurement or evaluation more frequently than annually to closely track program performance and to optimize the program delivery.

4.1 IPMVP Option

This protocol's recommended evaluation approach aligns best with International Performance Measurement and Verification Protocol (IPMVP) Option C, which recommends statistical analysis of data from utility meters for whole buildings or facilities to estimate savings. Option C is intended for projects with expected savings that are large relative to consumption. This protocol recommends regression analysis of residential customer consumption and statistical power analysis to determine the analysis sample size necessary to detect the expected savings.

4.2 Sample Design

Utilities should integrate the design of the analysis sample with program planning, because numerous considerations, including the size of the analysis sample, the method of recruiting customers to the program, and the type of randomized experiment, must be addressed before the program begins.

4.2.1 Sample Size

The analysis sample should be large enough to detect the minimum hypothesized program effect with desired probability.[28] If the sample is too small, evaluators risk being unable to detect the program's effect and wrongly accepting the hypothesis of no effect. Or there may be substantial uncertainty about the program's effect at the end of the study, and it may be necessary to repeat the study with a larger sample. On the other hand, if the sample size is too large, researchers may risk wasting scarce program resources.[29]

To determine the minimum number of subjects required and the number of subjects to be assigned to the treatment and control groups, researchers should employ a statistical power analysis. Statistical power is the likelihood of detecting a program impact of minimum size (the minimum detectable effect). Typically, researchers design studies to achieve statistical power of 80% or 90%. A study with 80% statistical power has an 80% probability of detecting the hypothesized treatment effect.

Statistical power analysis can be conducted in two ways. First, if data on consumption or another outcome of interest before treatment are available for the study population, researchers can use simulation to estimate the probability of detecting an effect of a certain size (for example, 1%) for possible treatment and control groups sizes, N_T and N_C.

Simulation follows these steps:

1. Researchers should divide the pre-treatment sample period into two parts, corresponding to a simulation pre-treatment and post-treatment period. For example, an evaluator with monthly billing consumption data for 24 pre-treatment months could divide the pre-

[28] A program can consist of a collection of randomized cohorts or waves in which the treatment effect of interest is at the program level and not at the level of individual cohorts. In this case, power calculations and tests of statistical significance can be applied to the collection of cohorts. Examples of this design include behavioral programs that consist of several waves launched over time or rolling enrollment waves.

[29] The utility may also base the number of subjects in the treatment group on the total savings it desires to achieve.

treatment period into months one to 12 and months 13 to 24 and designate the first 12 months as the simulation pre-treatment period.

2. From the eligible program population, researchers should randomly assign N_T subjects to the treatment group and N_C subjects to the control group.
3. Researchers should decide upon the minimum detectable treatment effect (for example, 2 kWh/period/subject), and a distribution of treatment effects (for example, normal distribution with mean 2 and standard deviation 1). For each treatment customer, the researcher should apply a treatment effect, taken randomly from the distribution of treatment effects, during the simulation treatment period. (One could also assume the treatment effect is the same for all customers and merely apply the same effect to all households; however, the power calculation is likely to underestimate the number of households needed because it assumes zero variance for the treatment effect.).
4. Researchers should randomly sample with replacement N_T customers from the treatment group and N_C subjects from the control group.
5. Researchers should estimate the program treatment effect for the sample only using data from the simulation pre-treatment and simulation post-treatment periods and retain the estimate.
6. Researchers should repeat steps 4 and 5 many times (for example, >250), and calculate the percentage of iterations that the estimated treatment effect was greater than zero. This is the statistical power of the study, the probability of detecting savings of x with treatment group size N_T and control group size N_C.

It is important that the estimation method used in the statistical power simulation adhere as closely as possible to the method evaluators plan to use for the actual savings estimation. Otherwise, the statistical power analysis may be misleading about the likelihood of detecting the savings.

The second approach to calculating statistical power uses analytic formulas. Researchers employing panel data methods and using statistical power formulas are advised to use the formulas in Burlig et al. (2017). Though more demanding to implement than those in Frison and Pocock (1992), the statistical power formulas in Burlig et al. (2017) are more accurate because they account for both intra-cluster correlations and arbitrary serial correlations of customer consumption over time. The required inputs for the power calculation are:

- The minimum detectable treatment effect
- The coefficient of variation of energy use, taken from a sample of customers
- The specific analysis approach to be used (for example, simple differences of means or a repeated measure analysis)
- The numbers of pre-treatment and post-treatment observations per subject
- The tolerances for Type I and Type II statistical errors (as discussed in Section 3.3)
- The intra-cluster correlation of an individual subject's energy use or error term covariances for pre-treatment and post-treatment periods and between periods.

Many statistical software, including SAS, STATA, and R, include packages for performing statistical power analyses.

Researchers conducting statistical power analyses should keep in mind the following:

- For a given program population, statistical power will be maximized if 50% of subjects are assigned to the treatment group and 50% are assigned to the control group. However, especially for large programs, researchers may obtain acceptable levels of statistical power with unbalanced treatment and control groups. The principal benefit of a smaller control group is that more customers are available to participate in the program.
- If the BB program will operate for more than several months and repeated measurements are planned, researchers should adjust the required sample sizes to account for attrition, the loss of some subjects from the analysis sample because of account closures or withdrawal from the study.

4.2.2 Random Assignment to Treatment and Control Groups by Independent Third Party

After determining the appropriate sizes of the treatment and control group samples, researchers should randomly assign subjects to the treatment and control groups. For the study to have maximum credibility and acceptance, this protocol recommends that an independent and experienced third party such as an independent evaluator perform the randomization. If there is a significant risk that the random assignment will result in unbalanced treatment and control groups, this protocol recommends that evaluators first stratify the study population by pretreatment energy use and then randomly assign subjects in each stratum to treatment and control groups. Stratifying the sample will increase the likelihood that treatment and control group subjects have similar pretreatment means and variances.[30]

This protocol also recommends that the unit of analysis (for example, a household) should be the basis for random assignment to treatment or control group. For example, in an analysis of individual customer consumption, it is better to randomly assign individual customers instead of all customers in the same neighborhood (for example, in a zip code or census block) to receive the treatment. However, for some BB programs, it may not be feasible to randomize the unit of analysis. For example, in some multifamily housing BB programs, the unit of analysis may be individual customers but all customers in the same multifamily building may receive the treatment. In this case, it will be necessary to randomly assign multifamily buildings to the treatment or control group. In this case, researchers will need to account for correlations in consumption between customers in the same housing units.

Although this protocol recommends that an independent and experienced third party perform the random assignment, circumstances sometimes make this impossible. In such cases, a third-party evaluator should certify that the assignment of treatment and control group subjects was done correctly and did not introduce bias into the selection process.

4.2.3 Equivalency Check

The third party performing the random assignment must verify that the characteristics of subjects in the treatment group, including pretreatment energy use, are balanced with those in the control

[30] Shadish et al. (2002) discuss the benefits of stratified random assignment. Bruhn and McKenzie (2009) compare stratified random assignment and re-randomization methods and finds that stratification is superior.

group. If subjects in the groups are not equivalent overall, the energy savings estimates may be biased.

To verify the equivalence of energy consumption, this protocol recommends that the third-party test for differences between treatment and control group subjects in both the mean pretreatment period energy consumption and in the distribution of pretreatment energy consumption. Evaluators should attempt to verify equivalence of energy consumption using the same frequency of data to be used in the savings analysis. For example, evaluators should use hour interval consumption data to verify equivalence if the study objective is to estimate peak hour energy savings. Evaluators should also test for differences in other available covariates, such as home floor area and heating fuel type. Evaluators can use t-tests or regression to conduct the tests. Section 4.4 describes the use of regression for verifying the equivalence of the two groups.

If significant differences are found, the third party should consider performing the random assignment again. Ideally, random assignment should not result in any differences; however, differences occasionally appear, and it is better to redo the random assignment than to proceed with unbalanced treatment and control groups, which may lead to biased savings estimates. As noted in Section 4.2.2, stratifying the study population by pretreatment energy use will increase the probability that the groups are balanced.

If the evaluator is not the third party who performed the random assignment, the evaluator should also perform an equivalency check. The evaluator may be able to use statistical methods to control for differences in pretreatment energy use that are found after the program is underway.[31]

4.3 Data Requirements and Collection

4.3.1 Energy Use Data

Estimating BB program impacts using a field experiment requires collecting energy use data from subjects in the analysis sample. This protocol recommends that evaluators collect multiple energy use measurements for each sampled unit for the periods before and during the treatment.[32]

These data are known as a panel. Panels can consist of multiple hourly, daily, or monthly energy use observations for each sampled unit. In this protocol, a panel refers to a dataset that includes energy measurements for each sampled unit either for the pretreatment and treatment periods or for the treatment period only. The time period for panel data collection will depend on the program timeline, the frequency of the energy use data, and the amount of data collected.

Panel data have several advantages for use in measuring BB program savings:

- **Relative ease of collection.** Collecting multiple energy use measurements for each sampled unit from utility billing systems is usually easy and inexpensive.

[31] If energy use data are available for the periods before and during the treatment, it is possible to control for time-invariant differences between sampled treatment and control group subjects using subject fixed effects.

[32] A single measurement of energy use for each sampled unit during the treatment period also results in an unbiased estimate of program savings. The statistical significance of the savings estimate depends on the variation of the true but unknown savings and the number of sampled units.

- **Can estimate savings during specific times.** If the panel collects enough energy use observations per sampled unit, estimating savings at specific times during the treatment period may be possible. For example, hourly energy use data may enable the estimation of precise savings during utility system peak hours. Monthly energy use data may enable the development of precise savings estimates for each month of the year.
- **Savings estimates are more precise.** Evaluators can more precisely estimate energy savings with a panel, because they may be able to control for the time-invariant differences in energy use between subjects that contribute to the variance of energy use.
- **Allows for smaller analysis samples.** All else being equal, fewer units are required to detect a minimum level of savings in a panel study than in a cross-section analysis. Thus, collecting panel data may enable studies with smaller analysis samples and data collection costs.

Using panel data has some disadvantages relative to a single measurement per household in a cross-sectional analysis. First, evaluators must correctly cluster the standard errors within each household or unit (as described in the following section). Second, panel data require statistical software to analyze, whereas estimating savings using single measurements in a basic spreadsheet software program may be possible.

This protocol also recommends that evaluators collect energy use data for the duration of the treatment to ensure they can observe the treatment effect for the entire study period. Ideally, an energy efficiency BB program lasts for a year or more because the energy end uses affected by BB programs vary seasonally. For example, these programs may influence weather-sensitive energy uses, such as space heating or cooling, so collecting less than 1 year of data to reflect every season may yield incomplete results.

Collecting data for an entire year may be impossible because some BB programs do not last that long. For these programs, only an unbiased estimate of savings for the time period of analysis may be obtained. Evaluators should exercise caution in extrapolating those estimates to seasons or months outside the analysis period, especially if the BB program affected weather-sensitive or seasonally varying end uses of energy.

4.3.2 Makeup of Analysis Sample

Evaluators must collect energy use measurements for every household or unit that is initially assigned to a control or treatment group, whether or not the household or unit later opts out. Not collecting energy use data for households initially placed in a treatment group but that then opts out results in imbalanced treatment and control groups and a biased savings estimate.

4.3.3 Other Data Requirements

Program information about each participant must also be collected. These data must include whether the subject was assigned to the treatment or control group, when the treatments were administered, and if and when the subject opted out.

Temperature and other weather data may also be useful but are usually not necessary. Often researchers can use dummy variables for individual time periods to account for the effect of weather on household energy consumption. If weather data will be collected, evaluators should obtain them from the weather station nearest to each household.

4.3.4 Data Collection Method

Energy use measurements used in the savings estimation should be collected directly from the utility, not from the program implementer, at the end of the program evaluation period. Depending on the program type, utility billing system, and evaluation objectives, the data frequency can be at 15-minute, 1-hour, daily, or monthly intervals.

4.4 Analysis Methods

This protocol recommends using panel regression analysis to estimate savings from BB field experiments where subjects were randomly assigned to either treatment or control groups. Evaluators typically prefer regression analysis to simply calculating differences in unconditional mean energy use, because it generally results in more precise savings estimates. A significant benefit of randomized field experiments is that regression-based savings estimates are usually quite insensitive to the type of model specification.

Section 4.3.1 addresses issues in panel regression estimation of BB program savings, including model specification and estimation, standard errors estimation, robustness checks, and savings estimation. It illustrates some specifications as well as the application of energy-savings estimation.

4.4.1 Panel Regression Analysis

In panel regressions, the dependent variable is usually the energy use of a subject (a home, apartment, or dormitory) per unit of time such a month, day, or hour. The right side of the equation includes an independent variable to indicate whether the subject was assigned to the treatment or control group. This variable can enter the model singularly or be interacted with another independent variable, depending on the analysis goals and the availability of energy use data from before treatment. The coefficient on the term with the treatment indicator is the energy savings per subject per unit of time. DiD models of energy savings must also include an indicator for whether the period occurred before or during the treatment period.

Many panel regressions also include fixed effects. Subject fixed effects capture unobservable energy use specific to a subject that does not vary over time. For example, home fixed effects may capture variation in energy use that is due to differences such as home sizes or makeup of a home's appliance stock. Time-period fixed effects capture unobservable energy use specific to a time period that does not vary between subjects. Including time or subject fixed effects in a regression of energy use of subjects randomly assigned to the treatment or control group will increase the precision but not the unbiasedness of the savings estimates.

Fixed effects can be incorporated into panel regression in several ways.

- Include a separate dummy variable or intercept for each subject in the model. The estimated coefficient on a subject's dummy variable represents the subject's time-invariant energy use. This approach, known as least squares dummy variables, may, however, not be practical for evaluations with a large number of subjects, because the model requires thousands of dummy variables that may overwhelm available computing resources.
- Apply the fixed-effect estimator, which requires transforming the dependent variable and all the independent variables by subtracting subject-specific means and then running

ordinary least squares (OLS) on the transformed data.[33] This approach is equivalent to least squares dummy variables.

- Estimate a first difference or annual difference of the model. Differencing removes the subject fixed effect and is equivalent to the dummy variable approach if the fixed-effects model is correctly specified.[34]

4.4.2 Panel Regression Model Specifications

This section outlines common regression approaches for estimating treatment effects from residential BB programs. Unless otherwise stated, assume that the BB program was implemented as a field experiment with an RCT or randomized encouragement design.

4.4.3 Simple Differences Regression Model of Energy Use

Consider a BB program in which the evaluator has energy use data for the treatment period only, and wishes to estimate the average energy savings per period from the treatment. Let $t = 1, 2, \ldots, T$, where t denotes the time periods during the treatment for which data are available,[35] and let $i = 1, 2, \ldots, N$, where i denotes the treatment and control group subjects. For simplicity, assume that all treated subjects started the treatment at the same time.

A basic specification to estimate the average energy savings per period from the treatment is:

Equation 1

$$y_{it} = \beta_0 + \beta_1 * Tr_i + \varepsilon_{it}$$

Where:

y_{it} = The metered energy use of subject i in period t.

β_0 = The average energy use per unit of time for subjects in the control group.

---

[33] Greene (2011) Chapter 11 provides more details.

[34] Standard econometric formulations assume that fixed effects account for unobservable factors that are correlated with one or more independent variables in the model. This correlation assumption distinguishes fixed-effects panel model estimation from other types of panel models. Fixed effects eliminate bias that would result from omitting unobserved time-invariant characteristics from the model. In general, fixed effects must be included to avoid omitted variable bias. In an RCT, however, fixed effects are unnecessary to the claim that the estimate of the treatment effect is unbiased because fixed effects are uncorrelated with the treatment by design. Although fixed effects regression is unnecessary, it will increase precision by reducing model variance.
Some evaluators may be tempted to choose to use random-effects estimation, which assumes time- or subject-invariant factors are uncorrelated with other variables in the model. However, fixed-effects estimation has important advantages over random-effects estimation: (1) it is robust to the omission of any time-invariant regressors. If the evaluator has doubts about whether the assumptions of the random-effects model are satisfied, the fixed-effects estimator is better; and (2) it yields consistent savings estimates when the assumptions of the random-effects model holds. The converse is not true, making the fixed-effects approach more robust.
Because weaker assumptions are required for the fixed-effects model to yield unbiased estimates, this protocol generally recommends the fixed-effects estimation approach. The remainder of this protocol presents panel regression models that satisfy the fixed-effects assumptions.

[35] For a treatment that is continuous, an example might be $t = 1$ on the first day that the treatment starts, $t = 2$ on the second day, etc.; for a treatment that occurs during certain days only (for example, a day when the utility's system peaks), an example might be $t = 1$ during the first critical event day, $t = 2$ during the second, etc.

β_1 = The average treatment effect of the program. The energy savings per subject per period equals $-\beta_1$.

Tr_i = An indicator for whether subject i received the treatment. The variable equals 1 for subjects in the treatment group and equals 0 for subjects in the control group.

ε_{it} = The model error term, representing random influences on the energy use of customer i in period t.

In this simple model, the error term ε_{it} is uncorrelated with Tr_i because subjects were randomly assigned to the treatment or control group. The OLS estimation of this model will result in an unbiased estimate of β_1. The standard errors should be clustered on the subject. [36]

This specification does not include subject fixed effects. Because the available energy use data apply to the treatment period only, the program treatment effect cannot be identified and subject fixed effects cannot be incorporated in the model. However, as previously noted, because of the random assignment of subjects to the treatment group, any time-invariant characteristics affecting energy use will be uncorrelated with the treatment, so omitting that type of fixed effects will not bias the savings estimates.

Using Equation 1, however, more precise estimates of savings could be obtained by replacing the coefficient β_0 with time-period fixed effects. The model thus captures more of the variation in energy use over time, resulting in greater precision in the estimate of savings. The interpretation of β_1, the average treatment effect per home per time period, is unchanged.

4.4.4 Simple Differences Regression Estimate of Heterogeneous Savings Impacts

Suppose that the evaluator still has energy use data that apply to the treatment period only, but wishes to obtain an estimate of savings from the treatment as a function of some exogenous variable such as preprogram energy use, temperature, home floor space, or pretreatment efficiency program participation (to determine, for example, whether high energy users save more or less energy than low energy users). If data for treatment and control group subjects on the exogenous variable of interest are available, the evaluator may be able to estimate the treatment effect as a function of this variable.

Let m_{ij} be an indicator that subject i belongs to a group j, j = 1, 2, …, J, where membership in group j is exogenous to receiving the treatment. Then the average treatment effect for subjects in group j can be estimated using the following regression equation:

Equation 2

$$y_{it} = \beta_0 + \Sigma^{J}_{j=1} \beta_{1j}*Tr_i*m_{ij} + \Sigma^{J-1}_{j=1} \gamma_j m_{ij} + \varepsilon_{it}$$

Where:

[36] Although the methods recommended in this protocol minimize the potential for violations of the assumptions of the classical linear regression model, evaluators should be aware of–and take steps to minimize—potential violations.

m_{ij} = An indicator for membership of subject i in group j. It equals 1 if customer i belongs to group j and equals 0, otherwise.

β_{1j} = The average treatment effect for subjects in group j. Energy savings per subject per period j equals $-\beta_{1j}$.

γ_j = The average energy use per period for subjects in group j, j = 1, 2,…J-1.

All of the other variables are defined as in Equation 1.

This specification includes a separate intercept for each group indicated by γ_j and the treatment indicator Tr_i interacted with each of the m_{ij} indicators. The coefficients on the interaction variables β_{1j} show average savings for group j relative to baseline average energy use for group j.

4.4.5 Simple Differences Regression Estimate of Savings During Each Time Period

To estimate the average energy savings from the treatment during each period, the evaluator can interact the treatment indicator with indicator variables for the time periods as in the following equation[37]:

Equation 3

$$y_{it} = \Sigma^{T}_{j=1} \beta_j Tr_i{*} d_{jt} + \Sigma^{T}_{j=1} \theta_j d_{jt} + \varepsilon_{it}$$

Where:

β_t = The average savings per subject specific to period t (for example, the average savings per subject during month 4 or during hour 6).

d_{jt} = An indicator variable for period j, j = 1, 2, …,T. d_{jt} equals 1 if j = t (that is, the period is the tth) and equals 0 if $j \neq t$ (that is, the period is not the tth).

θ_t = The average effect on consumption per subject specific to period t.

Equation 3 can be estimated by including a separate dummy variable and an interaction between that dummy variable and Tr_i for each time period t, where t = 1, 2, ..., T. When the time period is in months, the time-period variables are referred to as month-by-year fixed effects. The coefficient on the interaction variable for period t, β_t, is the average savings per subject for period t. Again, because ε_{it} is uncorrelated with the treatment after accounting for the average energy use in period t, the OLS estimation of Equation 3 (with standard errors clustered at the subject level) results in an unbiased estimate of the average treatment effect for each period.

Evaluators with smart meter data can use this specification to estimate BB program demand savings during specific hours of the analysis period. The coefficient β_t would indicate the demand savings from the treatment during hour *t*. Examples of research that estimates savings during hours of peak usage include Stewart (2013a) and Todd (2014).

[37] If the number of time periods is very large, the number of time period indicator variables in the regression may overwhelm the capabilities of the available statistical software. Another option for estimation is to transform the dependent variable and all of the independent variables by subtracting time period-specific means and then running the OLS on the transformed data.

4.4.6 Difference-in-Differences Regression Model of Energy Use

This section outlines a DiD approach to estimating savings from BB field experiments. This protocol recommends DiD estimation to the simple differences approach, but it requires information about the energy use of treatment and control group subjects during the pretreatment and treatment periods. These energy use data enable the evaluator to:

- Include subject fixed effects to account for differences between subjects in time-invariant energy use.
- Obtain more precise savings estimates.
- Test identifying assumptions of the model.

Assume there are N subjects and T +1 periods, T > 0, in the pretreatment period denoted by t = -T, -T+1, …, -1, 0, and T periods in the treatment period, denoted by t = 1, 2, …, T. A basic DiD panel regression with subject fixed effects could be specified as:

Equation 4

$$y_{it} = \alpha_i + \beta_1 P_t + \beta_2 P_t * Tr_i + \varepsilon_{it}$$

Where:

α_i = Unobservable, time-invariant energy use for subject i. These effects are controlled for with subject fixed effects.

β_1 = The average energy savings per subject during the treatment period that was not caused by the treatment.

P_t = An indicator variable for whether time period t occurs during the treatment. It equals 1 if treatment group subjects received the treatment during period t, and equals 0 otherwise.

β_2 = The average energy savings due to the treatment per subject per unit of time.

The model includes fixed effects to account for differences in average energy use between subjects. Including subject fixed effects would likely explain a significant amount of the variation in energy use between subjects and result in more precise savings estimates. The interaction of P_t and Tr_i equals one for subjects in the treatment group during periods when the treatment is in effect, and 0 for other periods and all control subjects.

Equation 4 is a DiD specification. For control group subject i, the expected energy use is α_i during the pretreatment period and $\alpha_i + \beta_1$ during the treatment period. The difference in expected energy use between pretreatment and treatment periods, also known as *naturally occurring savings*, is β_1. If that same subject i had been in the treatment group, the expected energy use would have been α_i during the pretreatment period and $\alpha_i + \beta_1 + \beta_2$ during the treatment period. The expected savings would have been $\beta_1 + \beta_2$, which is the sum of naturally occurring savings and savings from the BB program. Taking the difference yields β_2, a DiD estimate of program savings. The OLS estimation results in an unbiased estimate of β_2.

A more general form of Equation 4 would allow the treatment period to vary for each subject and substitute time-period fixed effects (such as a separate indicator variable for each day or month of the analysis period) for the stand-alone variable post-variable. This specification can be handy when subjects begin the treatment at different times such as with rolling program enrollments or if it is difficult to define when treatment would have begun for a control group subject.

Equation 5

$$y_{it} = \alpha_i + \tau_t + \beta_2 P_{it} * Tr_i + \varepsilon_{it}$$

Where:

τ_t = The time-period fixed effect (an unobservable that affects the consumption of all subjects during time period t). The time period effect can be estimated by including a separate dummy variable for each time period t, where t = -T, -T+1, …, -1, 0, 1, 2, ..., T.

P_{it} = An indicator variable for whether time period t occurs during the treatment for subject i. It equals 1 if treatment group subject i received the treatment during period t, and equals 0 otherwise.

As in Equation 4, the coefficient β_2 represents the average savings per customer per time period. The interpretations of the other variables and coefficients in the model remain unchanged.

4.4.7 DiD Estimate of Savings for Each Time Period

By respecifying Equation 4 with time-period fixed effects, savings can be estimated during each period and the identifying assumption tested to determine that assignment to the treatment was random. Consider the following DiD regression specification:

Equation 6

$$y_{it} = \alpha_i + \Sigma^{T}_{j=-T}\, \theta_j d_{jt} + \Sigma^{-1}_{j=-T}\, \beta_j\, Tr_i * d_{jt} + \Sigma^{T}_{j=1}\, \beta_j\, Tr_i * d_{jt} + \varepsilon_{it}$$

Savings in each period are estimated by including a separate dummy variable and an interaction between the dummy variable and Tr_i for each time period t, where t = -T, -T+1, …, -1, 0, 1, 2, ..., T. The coefficient on the interaction variable for period t, β_t^T, is the DiD savings for period t.

Unlike the simple differences regression model, this model yields an estimate of BB program savings during all periods except one, that is, t = 0, for a total of 2T-1 period savings estimates. Figure 4 shows an example of savings estimates obtained from such a model. The dotted lines show the 95% confidence interval for the savings estimates using standard errors clustered on utility customers.

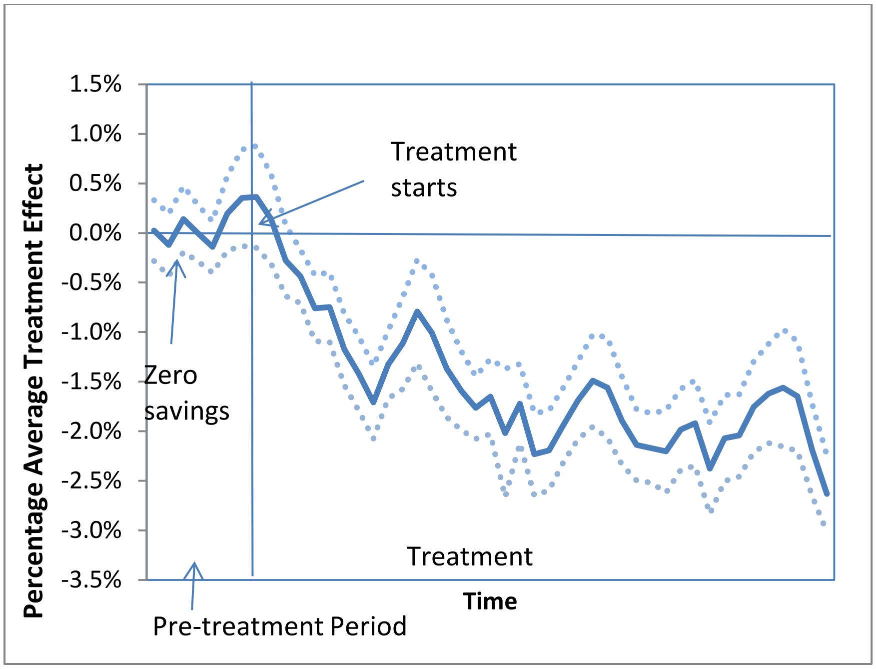


Figure 4. Example of DiD regression savings estimates

Estimates of pretreatment savings can be used to test the assumption of random assignment to the treatment. Before utilities administer the treatment, statistically significant differences in energy use between treatment and control group subjects should not be evident. BB program pretreatment saving estimates that were statistically different from zero would suggest a flaw in the experiment design. For example, an error in the randomization process may result in assignments of subjects to the treatment and control groups that were correlated with their energy use.

As with Equation 3, this specification can be used to estimate demand savings during specific hours. Energy use data for hours before the treatment are required, however.

4.4.8 Simple Differences Regression Model with Pre-Treatment Energy Consumption

In addition to estimating energy savings as a DiD, evaluators can estimate savings as a simple difference conditional on subject average pre-treatment energy consumption. This estimator, often referred to as "post-only," includes pre-treatment energy consumption as an independent variable in the regression to account for differences between subjects in their post-treatment consumption, serving a purpose similar to that of customer fixed effects in the DiD model.[38] However, many researchers favor the post-only estimator because it has smaller variance than the standard fixed effects, DiD estimator when energy consumption is uncorrelated or weakly correlated over time.[39]

[38] This model is also sometimes referred to as lagged dependent variable or post-period regression with pre-period controls.

[39] Some researchers refer to this model as a "post-only" model; however, this name is misleading because the model uses pre-treatment consumption as an explanatory variable. In a personal correspondence with the authors, Hunt Allcott, who introduced this method in evaluation of Home Energy Reports, points out that if seasonal effects are

Consider the following regression specification:

Equation 7

$$y_{it} = \tau_t + \beta_1 * Tr_{it} + \rho\overline{y_i^{pre}} + \varepsilon_{it}$$

Where:

τ_t = The time-period fixed effect (an unobservable that affects consumption of all subjects during time period t). The time period effect can be estimated by including a separate dummy variable for each time period t, where t = -T, -T+1, …, -1, 0, 1, 2, ..., T.

β_1 = Coefficient for the average treatment effect of the program. The energy savings per subject per period equals $-\beta_1$.

Tr_{it} = An indicator variable for whether subject i received the treatment in period t. The variable equals 1 for subjects who receive the treatment in period t and equals 0 otherwise.

ρ = Coefficient indicating the effect of average pre-treatment consumption on consumption during the treatment period.

$\overline{y_i^{pre}}$ = Average consumption during the pre-treatment period for subject i.

ε_{it} = The model error term, representing random influences on the energy use of customer i in period t.

With random assignment of subjects to treatment and control groups, the OLS estimation of Equation 6 is expected to produce an unbiased estimate of the average savings per subject per period.

Evaluators can estimate slightly different versions of this model:

- **Savings for each treatment period.** Evaluators can include a treatment indicator variable for each period instead of a treatment indicator variable for the entire treatment period. This specification will produce an estimate of average savings per subject for each treatment period.
- **Additional pre-treatment consumption control variables.** Instead of one pre-treatment consumption variable, evaluators can include multiple pre-treatment consumption variables, such as variables for different seasons or months of a year, days of week, or hours of the day.
- **Additional control variables.** Evaluators can add other variables such as weather to the model. The addition of such variables might help to improve the precision of the savings estimates.

being estimated, this model "has slightly smaller standard errors and can be better at addressing naturally occurring randomization imbalances that may result in the baseline pretreatment energy usage differing between the control and treatment group."

4.4.9 Randomized Encouragement Design

Some field experiments involve an RED in which subjects are only encouraged to accept a BB measure, in contrast to RCTs in which a program administers a BB intervention. This section outlines the types of regression models that are appropriate for REDs, how to interpret the coefficients, and how to estimate savings from RED programs.

Evaluators can apply the model specifications previously described for RCTs to REDs. The model coefficients and savings are interpreted differently, however, and an additional step is required to estimate average savings for subjects who accept the behavioral intervention. Treatment in an RED is defined as receiving encouragement to adopt the BB intervention, rather than actually receiving the intervention as with RCTs.

Consider a field experiment with an RED that has energy consumption data for treatment and control group subjects available for the pretreatment and treatment periods. Equations 1 through 4 can be used to estimate the treatment effect, or the average energy consumption effect on those receiving encouragement. The estimate captures savings from compliers only, because never takers never accept the intervention, and always takers would accept the intervention with or without encouragement.

To recover the LATE, the savings from subjects who accept the treatment because of the encouragement, scale the estimate of β_2 by the inverse of the difference between the percentage of subjects in the treatment group who accept the intervention and the percentage of subjects in the control group who accept the intervention (which is zero if control group subjects are prohibited from accepting the intervention). Estimate this as:

Equation 8

$$\beta_2 / (\pi_T - \pi_C)$$

Where:

π_T = The percentage of treatment group subjects who accept the intervention.

π_C = The percentage of control group subjects who accept the intervention.

A related approach for obtaining an estimate of savings for the BB intervention in a RED study is instrumental variables, two-stage least squares (IV-2SLS). This approach uses the random assignment of subjects to the treatment as an instrumental variable for the decision by encouraged customers to participate in the program. The instrumental variable provides the exogenous variation necessary to identify the effect of endogenous participation on energy consumption. Participation is endogenous because the encouraged customers' decisions to participate is not random and depends on unobserved characteristics that may be correlated with energy consumption. For encouragement to be a valid instrument, it must be that encouragement affects only energy consumption through its impact on BB program participation.

In the first stage, the evaluator regresses a binary program participation decision variable on an indicator for whether the customer was randomly assigned to receive encouragement and other exogenous independent variables from the second-stage energy consumption equation. The evaluator then uses the regression to predict the likelihood of participation for each subject and time period. In the second stage, the evaluator estimates the energy consumption equation, substituting the first-stage predicted likelihood of participation for the variable indicating actual program participation. The estimated coefficient on the predicted likelihood of participation is the LATE for the BB intervention.

For a detailed method of using an IV approach see Cappers et al. (2013) and for a real-world example of the IV-2SLS approach applied to a home weatherization program implemented as an RED, see Fowlie et al. (2010).

4.4.10 Models for Estimating Savings Persistence

A utility offering a residential BB program may want to know what happens to savings during the second or third year of the program or after treatment stops. There are two kinds of savings effects to measure: the effect of continuing the intervention on consumption is called *savings during treatment*, and the effect on consumption after discontinuing the intervention is called *post-treatment savings.* Recently, researchers have conducted analyses or meta-analyses of savings persistence for home energy reports programs (Allcott and Rogers 2014; Khawaja and Stewart 2014; Olig and Young 2016; Skumatz 2016).

Suppose a utility implemented a BB program as an RCT and wants to measure the persistence of savings after the BB intervention stops. The utility started the treatment in period $t = 1$ and administered it for t^* periods. Beginning in period $t = t^*+1$, the utility stopped administering the intervention for a random sample of treated subjects. Evaluators can estimate the average savings c for subjects who continue to receive the treatment (continuing treatment group) and for those who stopped receiving the treatment after period t^* (discontinued treatment group).

Assuming pretreatment energy use data are available, the following regression equation can be used to estimate *savings during treatment* and *post-treatment savings*:

Equation 9

$$kWh_{it} = \alpha_i + \tau_t + \beta_1 P_{1,t}{*}Tc_i + \beta_2 P_{1,t}{*}Td_i + \beta_3 P_{2,t}{*}Tc_i + \beta_4 P_{2,t}{*}Td_i + \varepsilon_{it}$$

Where:

τ_t = The time-period fixed effect (an unobservable that affects the consumption of all subjects during time period t). The time period effect can be estimated by including a separate dummy variable for each time period t, where $t = -T, -T+1, \ldots, -1, 0, 1, 2, \ldots, T$.

β_1 = The average energy savings per continuing subject caused by the treatment during periods $t = 1$ to $t = t^*$.

$P_{1,t}$ = An indicator variable for whether subjects in the continued and discontinued treatment groups received the treatment during period t. It equals 1 if period t occurs between periods $t = 1$ and $t = t^*$ and equals 0 otherwise.

Tc_i = An indicator for whether subject i is in the continuing treatment group. The variable equals 1 for subjects in the continuing treatment group and equals 0 for subjects not in the continuing treatment group.

β_2 = The average energy savings per discontinuing subject caused by the treatment during periods $t = 1$ to $t = t^*$.

Td_i = An indicator for whether subject i is in the discontinuing treatment group. The variable equals 1 for subjects in the discontinuing treatment group and equals 0 for subjects not in the discontinuing treatment group.

β_3 = The average energy savings from the treatment for subjects in the continuing treatment group when $t>t^*$.

$P_{2,t}$ = An indicator variable for whether continuing treatment group subjects received the treatment and discontinued treatment group subjects did not receive the treatment during period t. It equals 1 if period t occurs after $t = t^*$ and equals 0 otherwise.

β_4 = The average energy savings for subjects in the discontinued treatment group when $t>t^*$.

The OLS estimation of Equation 9 yields unbiased estimates of *savings during treatment* (β_3) and *post-treatment savings* (β_4) because original treatment group subjects were assigned randomly to the continuing and discontinued treatment groups. Evaluators can expect that $\beta_4 \geq \beta_3$, that is, the average savings of the continuing treatment group will be greater than that of the discontinued treatment group. To estimate savings decay after treatment stops, evaluators can take the difference between savings during treatment (β_2) and post-treatment savings (β_4) for subjects in the discontinued treatment group.

Evaluators can test the identifying assumption of random assignment to the discontinued treatment group by comparing the savings of continuing and discontinuing treatment group subject between period t = 1 and t*. If assignment was random, their savings during this period are expected to be equal.

4.4.11 Standard Errors

Panel data have multiple energy use observations for each subject; thus, the energy use data are very likely to exhibit within-subject correlations. Many factors affecting energy use persist over time, and the strength of within-subject correlations usually increases with the frequency of the data. When standard errors for panel regression model coefficients are calculated, these correlations must be accounted for. Failing to do so will lead to savings estimates with standard errors that are biased downward.

This protocol strongly recommends that evaluators estimate robust standard errors clustered on subjects (the randomized unit in field trials) to account for within-subject correlation. Most statistical software programs, including STATA, SAS, and R, have regression packages that output regression-clustered standard errors.

Clustered standard errors account for having less information about energy use in a panel with N subjects and T observations per subject than in a dataset with N*T independent observations. Because clustered standard errors account for these within-subject energy-use correlations, they are typically larger than OLS standard errors. When there is within-subject correlation, OLS standard errors are biased downward and overstate the statistical significance of the estimated regression coefficients.[40]

4.4.12 Opt-Out Subjects and Account Closures

Many BB programs allow subjects to opt out and stop receiving the treatment. This section addresses how evaluators should treat opt-out subjects in the analysis, as well as subjects whose billing accounts close during the analysis period.

As a general rule, evaluators should include all subjects initially assigned to the treatment and control groups in the savings analysis.[41] For example, evaluators should keep opt-out subjects in the analysis sample. Opt-out subjects may have different energy use characteristics than subjects who remain in the program, and dropping them from the analysis would result in nonequivalent treatment and control groups. To ensure the internal validity of the savings, opt-out subjects should be kept in the analysis sample.

Sometimes treatment or control group subjects close their billing accounts after the program starts. Account closures are usually unrelated to the BB program or savings; most are due to households changing residences. Subjects in the treatment group should experience account closures for the same reasons and at the same rates as subjects in the control group; evaluators can thus safely drop treatment and control group subjects who close their accounts from the analysis sample.

However, if savings are correlated with the probability of an account closure, it may be best to keep subjects with account closures in the analysis sample. For example, if young households, which are the most mobile and likely to close their accounts, are also most responsive to BB programs, dropping these households from the analysis would bias the savings estimates downward,[42] and evaluators should keep these households in the analysis.

If evaluators drop customers who close their accounts during the treatment from the regression estimation, they should still count the savings from these subjects for periods during the treatment before customers closed their accounts. To illustrate, when estimating savings for a 1-year BB program, evaluators can estimate the savings from subjects who closed their accounts and from those who did not as the weighted sum of the conditional average program treatment effects in each month:

Equation 10

$$\text{Savings} = \sum_{m=1}^{12} -\beta_m * \text{Days}_m * N_m$$

Where:

m = Indexes the months of the year

---

[40] Bertrand et al. (2004) show when DiD studies ignore serially correlated errors, the probability of finding significant effects when there are none (Type I error) increases significantly.

[41] This protocol urges evaluators not to arbitrarily drop outlier energy use observations from the analysis unless energy use was measured incorrectly. If an outlier is dropped from the analysis, the reasons for dropping the outlier and the effects of dropping it from the analysis on the savings estimates should be clearly documented. Evaluators should test the sensitivity of the results to dropping observations.

[42] See State and Local Efficiency Action Network (2012), p. 30.

$-\beta_m$ = The conditional average daily savings in month m (obtained from a regression equation that estimates the program treatment effect on energy use in each month)

$Days_m$ = The number of days in month m

N_m = The number of subjects with active accounts receiving the treatment in month m or in a previous month.

This approach assumes that savings in a given month for subjects who close their accounts are equal to savings of subjects whose accounts remain open.

4.5 Energy Efficiency Program Uplift and Double Counting of Savings

Many BB programs increase participation in other utility energy efficiency programs; this additional participation is known as *efficiency program uplift*. For example, many utilities encourage their energy report program recipients to participate in their other energy efficiency programs that provide cash rebates in exchange for adopting efficiency measures such as efficient furnaces, air conditioners, wall insulation, windows, and compact fluorescent lamps.

Quantifying the effects of BB programs on efficiency program participation is important for two reasons:

- Uplift can be an important effect of BB programs and a potential additional source of energy savings.
- Savings from efficiency program uplift could be double-counted if unaccounted for. That is, when a household participates in an efficiency program because of a BB program intervention, the utility may count the program savings twice: once in the regression-based estimate of BB program savings and again in the estimate of savings for the rebate program. To avoid double-counting savings, evaluators must estimate savings from program uplift and subtract these savings from the efficiency program portfolio savings.[43]

Estimating savings from BB program uplift with randomized experiments recommended in this protocol is conceptually straightforward. To illustrate, suppose that a utility markets energy efficiency Measure A to treatment and control group subjects identically through a separate rebate program. Subjects in the treatment group also receive behavioral messaging encouraging them to adopt efficiency measures, including Measure A. Because customers were randomly assigned to the treatment and control groups, the groups are expected to be equivalent except for the treated customers who received the behavior treatment. Therefore, evaluators can attribute any difference in the uptake of Measure A between the groups to the BB program.

[43] This protocol does not take a position on which program gets credit for the uplift. When a BB intervention causes participation in an energy efficiency program, we know that the program participation would not have occurred without the intervention. However, the amount of uplift caused by the BB intervention may depend on the dollar incentives provided by the efficiency program. For example, the BB program may produce greater lift in participation for a program incentive of $200 than $100. To determine the relationship between uplift and the incentive amount, it would be necessary to randomize the incentive amount and to study participation as a function of incentives and who receives the BB intervention.

Figure 5 illustrates this logic for calculating behavior program savings from efficiency program uplift. Behavior program savings from adoption of Measure A is the difference between the treatment group and the control group in savings from Measure A.

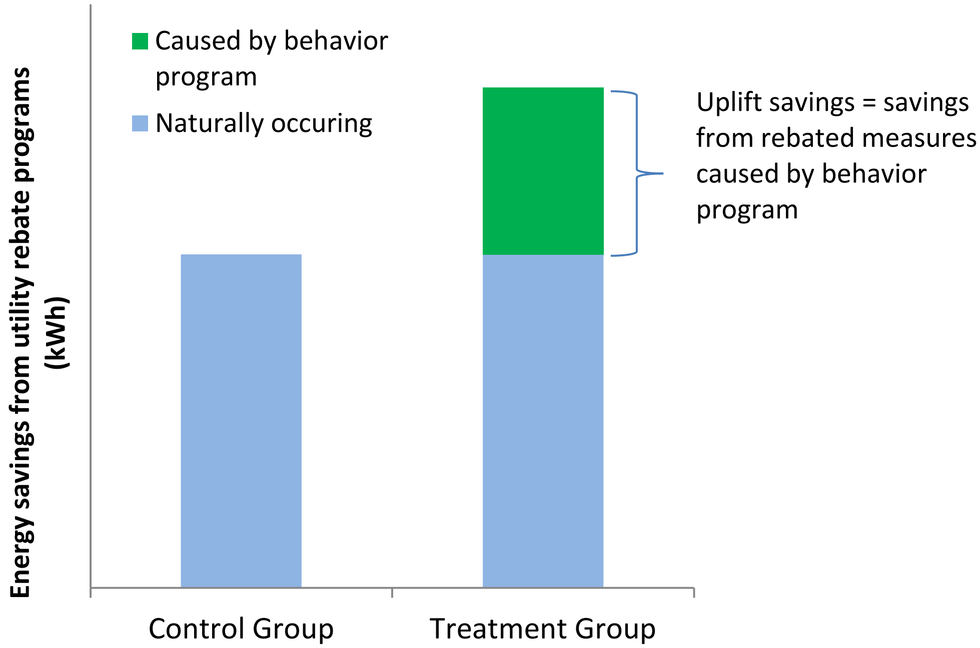


Figure 5. Calculation of double-counted savings

To estimate BB program savings from efficiency program uplift, evaluators should take the following steps:

1. Match the BB program treatment and control group subjects to the utility energy efficiency program tracking data.
2. Calculate the uplift savings per treatment group subject as the difference between treatment and control groups in average efficiency program savings per subject, where the savings are obtained from the utility tracking database of installed measures. (The averages should be calculated over all treatment group subjects and all control group subjects, not just those who participated in efficiency programs.)
3. Multiply the uplift savings per treatment group customer by the number of subjects who were in the treatment group to obtain the total uplift savings.

Evaluators can estimate BB program savings from efficiency program uplift for efficiency measures that the utility tracks at the customer level. Most measures for which utilities offer rebates—such as high-efficiency furnaces, windows, insulation, and air conditioners—fit this description.

Evaluators should be mindful of specific reporting conventions for efficiency program measures in utility tracking databases. For example, many jurisdictions require utilities to report weather-normalized and annualized measure savings, which do not reflect when measures were installed during the year or the actual weather conditions that affected savings. In contrast, the regression-based estimate of energy savings will reflect installation dates of measures and actual weather.

Evaluators should therefore adjust the annualized deemed savings in the program reporting database to account for when measures were installed during the year.

In addition, for BB programs running longer than one year, evaluators should account for the savings impacts of program uplift in previous years. Measures with a multiyear life installed in a previous program year will continue to save energy in subsequent years. Depending on the utility's conventions for reporting savings, it may be necessary to subtract savings from uplift in previous years from BB program savings estimate.

Estimating savings from program uplift for measures that the utility does not track at the customer level is more challenging. The most important such measures are high-efficiency lights such as compact fluorescent lamps and light-emitting diodes that are rebated through utility upstream programs. Most utilities provide incentives directly to retailers for purchasing these measures, and the retailers then pass on these price savings to utility customers in the form of retail discounts. Data on the purchases of rebated measures by treatment and control group subjects must be collected to estimate BB savings in upstream efficiency programs. Evaluators can use household surveys for this purpose.[44] However, because the difference in the number of purchased bulbs between treatment and control group subjects may be small, it may be necessary to survey a very large number of subjects to detect the BB program effect. Also, evaluators should adjust the lighting purchases impact estimates for in-service rates and the percentage of high efficiency lamps sold in the utility service area that received rebates. Evaluators should also be aware that some energy savings from purchasing compact fluorescent lamps or light-emitting diodes may be offset by reductions in the hours of use of those bulbs by treated customers.

[44] For an example of the approach required to estimate BB program savings from adoption of compact fluorescent lamps, see PG&E (2013).

5 Reporting

BB program evaluators should carefully document the research design, data collection and processing steps, analysis methods, and plan for calculating savings estimates. Specifically, evaluators should describe:

- The program implementation and the hypothesized effects of the behavioral intervention
- The experimental design, including the procedures for randomly assigning subjects to the treatment or control group
- The sample design and sampling process
- Processes for data collection and preparation for analysis, including all data cleaning steps
- Analysis methods, including the application of statistical or econometric models and key assumptions used to identify savings, including tests of those key identification assumptions
- Results of savings estimate, including point estimates of savings and standard errors and full results of regressions used to estimate savings.

A good rule-of-thumb is that evaluators should report enough detail such that a different evaluator could replicate the study with the same data. Every detail does not have to be provided in the body of the report; many of the data collection and savings estimation details can be provided in a technical appendix.

6 Looking Forward

Evaluators and program administrators should employ randomized experiments for evaluating BB programs whenever possible. However, some BB programs may be difficult or costly to evaluate using randomized experiments. In these cases, evaluators must employ quasi-experiments that rely on random but uncontrolled variation in who participates.

An important question concerns the accuracy of quasi-experimental methods such as propensity-score matching, regression discontinuity, and DiD estimation for evaluating BB programs. Evaluators of BB programs have employed and will continue to employ these methods. Although this protocol has cited several studies comparing the accuracy of randomized experiments and quasi-experiments, more research will be needed to draw firm conclusions about the accuracy of quasi-experiments.

Depending on the outcome of this research and acceptance by regulators and program administrators of savings estimates from quasi-experiments, evaluators could give consideration to updating this protocol to include quasi-experimental methods.

7 References

Allcott, H. (2015). "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics* 130 (3), pp. 1117-1165.

Allcott, H. (2011). "Social Norms and Energy Conservation." *Journal of Public Economics*, 95(2), pp. 1082-1095.

Allcott, H.; Rodgers, T. (2014). "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." *American Economic Review* 104 (10), pp. 3003-3037.

Baylis, P.; Cappers, P.; Jin, L.; Spurlock, A.; Todd, A. (2016). "Go for the Silver? Evidence from field studies quantifying the difference in evaluation results between "gold standard" randomized controlled trial methods versus quasi-experimental methods." ACEEE Summer Study on Energy Efficiency in Buildings 2016.

Bertrand, M.; Duflo, E.; Mullainathan, S. (2004). "How Much Should We Trust Difference-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (1), 249–275.

Brandon, A.; Ferraro, P. J.; List, J. A.; Metcalfe, R. D.; Price, M. K.; Rundhammer, F. (2017). *Do the Effects of Social Nudges Persist? Theory and Evidence from 38 Natural Field Experiments.* National Bureau of Economic Research working paper 23277. Available at: http://www.nber.org/papers/w23277

Bruhn, M., McKenzie D. (2009). "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1(4): 200-232.

Burlig, F., Preonas, L., Woerman, M. (2017). *Panel Data and Experimental Design. Energy Institute at Haas working paper.* Available at: https://ei.haas.berkeley.edu/research/papers/WP277.pdf

Cappers, P.; Todd, A.; Perry, M.; Neenan, B.; Boisvert, R. (2013). *Quantifying the Impacts of Time-based Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies: Protocols and Guidelines.* Lawrence Berkeley National Laboratory, Berkeley, CA and the Electric Power Research Institute, Palo Alto, CA. LBNL-6301E. Available at: https://emp.lbl.gov/sites/default/files/lbnl-6301e.pdf

Costa, D.L.; Kahn, M.E. (2010). *Energy Conservation "Nudges" and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment.* NBER Working Paper 15939. Available at: http://www.nber.org/papers/w15939

Consortium for Energy Efficiency Database (2013). Available at: http://library.cee1.org/content/2013-behavior-program-summary-public-version

Davis, M. (2011). *Behavior and Energy Savings: Evidence from a Series of Experimental Interventions.* Environmental Defense Fund Report.

EPRI (2010). *Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols.* Electric Power Research Institute: Palo Alto, CA. 1020855.

Fowlie, M. (2010). *U.S. Department of Energy Smart Grid Investment Grant Technical Advisory Group Guidance Document #7, Topic: Design and Implementation of Program Evaluations that Utilize Randomized Experimental Approaches.* November 8, 2010.

Frison, L.; Pocock, S. J. (1992). "Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design." *Statistics in Medicine* 11.13 (1992): pp. 1685-704.

Greene, W. (2011). *Econometric Analysis.* New Jersey: Prentice Hall.

Harding, M.; Hsiaw, A. (2012). *Goal Setting and Energy Efficiency.* Stanford University working paper.

Ignelzi, P.; Peters, J.; Dethman, L.; Randazzo, K.; Lutzenhiser, L. (2013). *Paving the Way for a Richer Mix of Residential Behavior Programs.* Prepared for Enernoc Utility Solutions. CALMAC Study SCE0334.01.

Khawaja, M.S., Stewart, J. (2015). *Long-Run Savings and Cost Effectiveness of Home Energy Report Programs.* Cadmus Research Report. Available at: http://www.cadmusgroup.com/wp-content/uploads/2014/11/Cadmus_Home_Energy_Reports_Winter2014.pdf

List, J.A. (2011). "Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off." *Journal of Economic Perspectives* (25), pp. 3–16.

List, J.A., Sadoff, S.; Wagner, M. (2010). *So You Want to Run an Experiment, Now What? Some Simple Rules of Thumb for Optimal Experimental Design.* National Bureau of Economic Research Paper 15701.

Mazur-Strommen, S.; Farley, K. (2013). *ACEEE Field Guide to Utility-Run Behavior Programs.* American Council for an Energy Efficient Economy, Report Number B132.

Minnesota Department of Commerce, Division of Energy Resources (2015). *Energy Efficiency Behavioral Programs: Literature Review, Benchmarking Analysis, and Evaluation Guidelines.* Prepared by Illume Advising with subcontractors Indica Consulting and Dr. Edward Vine. Available at: http://mn.gov/commerce-stat/pdfs/card-report-energy-efficiency-behavorial-prog.pdf

National Renewable Energy Laboratory. (2017). *Strategic Energy Management Protocol: The Uniform Methods Project.* Prepared by James Stewart. Available at: https://www.nrel.gov/docs/fy17osti/68316.pdf

Olig, C.; Young, E. 2016. *ComEd Home Energy Report Program Decay Rate and Persistence Study – Year Two.* Navigant.

PG&E (2013). *2012 Load Impact Evaluation for Pacific Gas and Electric Company's SmartAC Program*. Prepared by Freeman, Sullivan & Co.

PSE (2012). *Home Energy Reports Program: Three Year Impact Behavioral and Process Evaluation.* Puget Sound Energy. Prepared by DNV GL.

Rosenberg, M.; Kennedy Agnew, G.; Gaffney, K. (2013). *Causality, Sustainability, and Scalability – What We Still Do and Do Not Know about the Impacts of Comparative Feedback Programs.* Paper prepared for 2013 International Energy Program Evaluation Conference, Chicago.

SEE Action (2012). *Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations.* State and Local Energy Efficiency Action Network. Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman, Lawrence Berkeley National Laboratory. http://behavioranalytics.lbl.gov. Available at: https://www4.eere.energy.gov/seeaction/system/files/documents/emv_behaviorbased_eeprograms.pdf

Seelig, M.J. (2013) "Business Energy Reports" presentation at BECC 2013. Available at: http://beccconference.org/wp-content/uploads/2013/12/BECC_PGE_BER_11-19-13_seelig-.pdf

Shadish, W.R., Cook, T.D.; Campbell, D.T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Belmont, CA: Wadsworth Cengage Learning.

Skumatz, L. (2016). Persistence of Behavioral Programs: New Information and Implications for Program Optimization. *The Electricity Journal* 29 (5): pp. 27-32.

SMUD (2013). *SmartPricing Options Interim Evaluation.* Prepared for the U.S. Department of Energy and Lawrence Berkeley National Laboratory by Sacramento Municipal Utility District and Freeman, Sullivan, & Co.

SMUD (2011). *Sacramento Municipal Utility District Home Energy Report Program Impact and Persistence Evaluation Report, Years 2008-2011.* Prepared by May Wu, Integral Analytics.

Stewart, J. (2013a). *Peak-Coincident Demand Savings from Behavior-Based Programs: Evidence from PPL Electric's Behavior and Education Program.* UC Berkeley: Behavior, Energy and Climate Change Conference. Available at: http://escholarship.org/uc/item/3cc9b30t.

Stewart, J. (2013b). "Energy Savings from Business Energy Feedback." Presentation at BECC 2013. Available at: http://beccconference.org/wp-content/uploads/2015/10/presentation_stewart.pdf.

Todd, A. (2014). *Insights from Smart Meters: The Potential for Peak-Hour Savings from Behavior-Based Programs.* Lawrence Berkeley National Laboratory. LBNL Paper LBNL-6598E. Available at: http://escholarship.org/uc/item/2nv5q42n

Chapter 18: Variable Frequency Drive Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in August 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in November 2014.

Jeff Romberger
SBW Consulting, Inc.
Bellevue, Washington

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68574
August 2017

Contract No. DE-AC36-08GO28308

Chapter 18: Variable Frequency Drive Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in August 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in November 2014.

Jeff Romberger
SBW Consulting, Inc.
Bellevue, Washington

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68574
August 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at https://energy.gov/sites/prod/files/2015/02/f19/UMPIntro1.pdf.

Acknowledgments

The chapter author wishes to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Mike Rufo of Itron
- David Jacobson of Jacobson Energy Research LLC.

Suggested Citation

Romberger, Jeff. 2017. *Chapter 18: Variable Frequency Drive Evaluation Protocol The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68574. http://www.nrel.gov/docs/fy17osti/68574.pdf.

Acronyms

ASD	Adjustable-speed drive
BAS	building automation system
CV	constant volume
HVAC	heating, ventilating, and air conditioning
IPMVP	International Performance Measurement and Verification Protocol
M&V	measurement and verification
OAT	outside air temperature
RMS	root mean square
RTF	Regional Technical Forum
TMY	typical meteorological year
VAV	variable air volume
VFD	variable frequency drive
VSD	variable-speed drive

Table of Contents

List of Tables

1 Measure Description

An adjustable-speed drive (ASD) includes all devices that vary the speed of a rotating load, including those that vary the motor speed and linkage devices that allow constant motor speed while varying the load speed. The Variable Frequency Drive Evaluation Protocol presented here addresses evaluation issues for variable-frequency drives (VFDs) installed on commercial and industrial motor-driven centrifugal fans and pumps for which torque varies with speed.[1] Constant torque load applications, such as those for positive displacement pumps, are not covered by this protocol. Other ASD devices, such as magnetic drive, eddy current drives, variable belt sheave drives, or direct current motor variable voltage drives, are also not addressed. The VFD is by far the most common type of ASD hardware. With VFD speed control on a centrifugal fan or pump motor, energy use follows the affinity laws, which state that the motor electricity demand is a cubic relationship to speed under ideal conditions. Therefore, if the motor runs at 75% speed, the motor demand will ideally be reduced to 42% of full load power; however, with other losses it is about 49% of full load power.

VFDs are commonly used on other motor-driven equipment such as air compressors, refrigeration compressors, vacuum pumps, and high-pressure blowers. These devices are typically positive displacement machines and are not included under this protocol, but in some cases will be addressed in protocols that are specific to them.

This protocol is also not intended to address conditions where there is significant interaction with other end uses, such as heating or cooling. For example, VFDs on refrigeration evaporator fans are not addressed because the fans significantly impact refrigeration load. VFDs on cooling tower fans are not addressed because the fans are often combined with condenser water temperature control, which impacts the chiller energy use. Conversion of constant volume (CV) heating, ventilation, and air conditioning (HVAC) systems to variable air volume (VAV) systems can have significant impacts on heating and cooling loads.

In some cases no interaction may occur, such as CV-to-VAV conversion of parking garage ventilation fans. These may be applicable to this protocol because no interaction with other end uses takes place. Other cases may be considered if the interaction is expected to be small compared to the fan motor energy savings.

[1] As discussed in *Considering Resource Constraints* in the Introduction of this report, small utilities (as defined under the U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

2 Application Conditions of Protocol

Market transformation is occurring for VFDs on HVAC fans and pumps for new construction. Building codes vary by jurisdiction, but some require VFDs on all HVAC fans and pumps for a certain size, such as the California Title 24 building code, which requires VFDs on all HVAC fans and pumps greater than 10 horsepower. Some jurisdictions continue to use building codes that do not yet address VFDs as a requirement.

In general, building codes do not address VFD requirements for industrial process fans and pumps. Retrofit of existing HVAC pumps and fans with VFDs remains a common application.

Generally, VFDs for HVAC applications tend not to save electricity during system peak hours, because that is often the time of peak HVAC fan and pump demand. The VFD electricity demand will be greater during peak use (if the drive is operating at full speed or faster during peak periods), because the VFD is typically about 97% efficient at full speed. Given that this depends on the peak savings definition and defined peak period, some jurisdictions may have peak savings. To determine if peak savings may occur, care should be taken to understand the load profile during the peak period of the motor on which VFD is installed. There are some specific cases where savings do occur during peak periods, such as a chilled water pump that has a throttling valve that is opened fully after the VFD is installed, which allows the VFD to operate at less than full speed even at peak flow.

Energy efficiency programs encourage the use of VFDs (as retrofits and in new construction) on fans and pumps that serve loads that vary over time, even if the local energy code does not require them. Three mechanisms for program delivery are commonly used across the United States.

- **Prescriptive.** This approach provides an incentive and deems energy savings based on the installed motor horsepower[2]. The incentive and energy savings may also vary based on the building type and fan/pump application where the VFD is installed, because the energy savings will typically vary for different installations (e.g., hospitals versus office buildings). In other cases, incentives and deemed energy savings may be designed based on horsepower and the annual operating hours of the equipment.
- **Standard calculator.** This approach provides an incentive for the VFD based on the expected annual energy savings, in kilowatt-hours, estimated using a standard calculation tool. The standard calculator usually incorporates the default performance curves used in DOE-2.X, based hourly simulation models such as eQUEST[3] or EnergyPlus.[4] Standard

[2] A recent study (Cadmus Loadshape, 2014) has derived average kWh and kW savings values for VFD installation on HVAC fans and pumps based on direct and long-term measurements of nearly 400 VSD installations accounting for the diversity of motor sizes, building types, HVAC loads, operating strategies, and seasonal differences across the northeast. These values will be useful for program implementers to deem VFD energy savings for future programs. The report makes several recommendations to implementers to maximize program effectiveness as well as important factors that evaluators should consider (consistent with this protocol).

[3] The eQUEST building energy simulation software is supported as a part of the Energy Design Resources program, which is funded by California utility customers and is available with documentation at the website: http://www.doe2.com/equest/.

[4] The EnergyPlus building energy simulation software is provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy. The software with documentation is available at the website: http://apps1.eere.energy.gov/buildings/energyplus/.

assumptions for operating hours are specified for the annual time that the VFD will operate at various speeds. These calculations may be customized if additional information is available, such as pre metering used to develop a flow profile. The baseline performance will also use the appropriate default curve for the baseline condition, such as for a fan with outlet dampers.

- **Custom.** This approach also provides an incentive for the VFD based on the expected annual energy savings (in kilowatt-hours), but savings are calculated using a custom calculation tool. This may be the result of a complete hourly building model that is developed using a program such as eQUEST or EnergyPlus. This custom calculation approach is more common for facilities that are applying incentives for a variety of measures in a building. Other calculation approaches may be used, such as developing a bin model for the HVAC systems in a building or for an industrial process, and may include metering data. Custom programs may require measurement and verification (M&V) after VFD installation to verify energy savings and determine incentive amounts based on actual equipment performance.

3 Savings Calculations

This section presents a high-level gross energy savings equation that applies to all VFD measures.[5] Detailed direction on how to apply this equation is presented under the Measurement and Verification Plan section of this protocol. Two approaches (performance curve and default curve methods) are presented based on the specifics of the pre- and post-implementation operating conditions for the specific application.

Energy savings should be determined using the following general equation (EVO 2012):

$$\text{Gross Energy Savings} = (\text{Baseline Energy} - \text{Reporting Period Energy}) \pm \text{Routine Adjustments} \tag{1}$$

Where,

Gross Energy Savings	=	Estimated typical annual energy consumption savings
Baseline Energy	=	Pre-implementation annual consumption
Reporting Period Energy	=	Post-implementation annual consumption
Routine Adjustments	=	Adjustments made to account for routine changes to independent variables (variables that drive energy consumption) that are not caused by the installation of the VFD. Savings should be normalized to typical meteorological year (TMY) weather data (preferably TMY3 data) as well as other significant independent variables (e.g., occupancy schedules, production data), if applicable. If first year energy savings is desired, then savings should be normalized to the actual weather for the 12-month period following commissioning of the new controls.

[5] This protocol focuses on gross energy savings and does not include other parameter assessments, such as net-to-gross, peak coincidence factors, or cost-effectiveness.

4 Measurement and Verification Plan

This section contains two approaches for determining VFD energy savings—the Performance Curve Method and the Default Curve Method—and guidance on how and when to use each. Both methods use post-installation metered data. Neither requires pre-installation metered data, but use of such data will improve reliability of savings estimates. The Performance Curve Method is more reliable, using the performance curves specific to the fan or pump on which the VFD is installed. This method is preferred whenever it is applicable. However, this method applies to a relatively narrow set of eligibility conditions.

The Default Curve Method uses default performance curves and is less precise, but it is applicable to a much wider set of conditions. This method could also be used for the conditions that specifically apply to the Performance Curve Method, but with less accuracy.

An alternative variation on the second method includes pre-installation (baseline) metering. This will improve the accuracy, but requires difficult and potentially expensive- baseline data and is therefore rarely used. This alternative requires metering both fan/pump motor power and air/water flow to develop the in-situ flow versus power relationship, which can be used instead of the default baseline curves. The flow trends can be expensive and difficult to accurately obtain and may delay project implementation. However, in some cases, flow measurements are available from the building automation system (BAS) and can be trended.

Timing of the post-installation evaluation is important. Customers may take a year or longer after installing the VFD to set up controls and fully commission the system. Performing evaluation activities within a year of installation will provide accurate first year results, but may not accurately reflect long term performance (Cadmus Loadshape, 2014).

4.1 Performance Curve Method

This method is consistent with International Performance Measurement and Verification Protocol (IPMVP) Option A (Isolation Retrofit, Key Parameter Measurement) (EVO 2012). The method incorporates kilowatt metering of the VFD installed on the fan/pump along with concurrent outside air temperature (OAT), if the load is temperature sensitive, and/or on another parameter such as production or schedule. The method estimates energy savings based on trended or logged power measurements, together with fan/pump performance specifications and site operating characteristics.

The basis for the calculation is that the post-installation power trend, combined with site system specification, allows the derivation of the post-installation flow trend. This flow trend is assumed to apply in the baseline case, where it is used to derive the baseline power trend from the pump performance curve. The OAT or other sensitive parameter is used to extrapolate individual savings values to an annual profile, which is summed to annual savings.

The Performance Curve Method has been developed as a standard protocol for fans (SBW Fan, 2012) and for pumps (SBW Pump, 2012), with provisional status, for the Regional Technical Forum (RTF). The RTF is an advisory committee established in 1999 to develop standards to verify and evaluate conservation savings in support of member utilities and other stakeholders in the Pacific Northwest Region. The specific protocol specifications are available for fans and pumps separately at the RTF website.[6] An Excel-based calculator has been developed for the

[6] http://rtf.nwcouncil.org/

pump application of this method and is also available at the RTF website. This section is a condensation of both the fan and pump RTF protocol documents. It lays out the data requirements as inputs to the calculator and describes the savings calculation methodology.

4.1.1 Eligible Projects

The primary eligibility requirement for the Performance Curve Method is that the system curve remains constant in post-installation operation. The system curve defines how pressure varies with flow due to the resistance to flow defined by the system configuration. More specifically, the following system requirements are necessary for this method to be accurately applied:

- Loads served must be similar pre- and post-installation, but airflow and water flow may vary by a different mechanism (outlet dampers and outlet throttling valve). If VFD controls are overridden and manually set, the only valid baseline would be if the flow were also manually set in a similar fashion. Controls in the baseline that change the fan or pump curve such as inlet vanes are not applicable.
- To ensure that the system curve remains constant in the post-installation period, operable dampers and throttling valves must be removed or disabled, and no dampers or throttling valves that change position during operation may remain. This is necessary to ensure that the system curve does not change, because any change would invalidate the methodology.
- The method may be applied to a single fan or pump. Multiple fans or pumps must be treated separately. Fans and pumps that are configured in parallel and that are controlled to operate at the same speed can be evaluated by this method, but fans in series would be excluded. Backup fans and pumps or multiple fans and pumps, where the same number operate in parallel, can also be evaluated with this method.
- Fan or pump motors retrofit with a VFD must be single-speed motors.
- Baseline control strategies that are not eligible for this method include, but are not limited to, variable-pitch blades, bypass, or cycling.

4.1.2 Data Collection Requirements

4.1.2.1 Fan/Pump and Motor Specifications

- **Fan/pump curve.** Data points from the manufacturer's performance curve include flow, pressure, and efficiency points from the appropriate fan/pump curve. The fan/pump curve must match the conditions at the site for impeller size and speed (revolutions per minute). Impeller size may be difficult to confirm, so as-built documents and maintenance records should be referenced to identify the original impeller size and to determine if the impeller has been trimmed.
- **Fan/pump motor hp.** These data are obtained from the motor nameplate.
- **Motor revolutions per minute.** These data are obtained from the motor nameplate.
- **Motor enclosure type.** This information is obtained from the motor nameplate.
- **Motor rated efficiency.** This information is obtained from the motor nameplate.

4.1.2.2 Fan/Pump Operations

- **Determinants of fan or pump speed.** Possible determinants are: (1) facility operation schedule, (2) OAT combined with operation schedule, and (3) production level.
- **Typical OAT.** Determine if the OAT is a significant determinant through discussion with the facility operator and consideration of the types of load being served by the fan or pump.
- **Facility operation schedule.** This information is obtained from the facility operator. If the facility's operation schedule has established different operation modes for the fan or pump unit (e.g., setback of flow during night and weekend hours), determine the period for each mode, defined as needed by hour of day, day of week, and season. This method requires that all schedule modes be metered.
- **Facility production level.** This information is obtained from the facility operator. If the facility's production schedule has established different operation modes for the fan or pump unit (e.g., two production lines for one work shift and one line for other shift), determine the period for each production level, defined as needed by hour of day, day of week, and season. This method requires that all production modes be metered.
- **Weather station.** If OAT is a significant determinant, identify the TMY weather station that is representative of the project site. If the weather station is not close by, adjustments may be needed (e.g., altitude differences).
- **Static head.** The head at zero flow—the net static head including elevation head for pumps—must be known. Site personnel should be able to provide this value or describe the means to acquire it.
- **System operating point.** This operating point (flow, pressure) must be with all dampers and valves removed or wide open. The point may be taken from the equipment schedule on the facility's mechanical plan. Alternatively, this value may be determined by: (1) inspecting facility control system trend logs of flow rate, if the system has a calibrated flow sensor and the log contains values at or near 100% speed; or (2) based on a pair of values (measured kilowatts and corresponding VFD speed).

4.1.2.3 Post-Period Measurements

- **True root mean square (RMS) power.** This protocol prefers a trend log of true poly-phase RMS power for the circuit powering the VFD and 15-minute intervals for the trend data. In general terms, a measurement period should be long enough to observe significant variation of dependent operating variables, such as OAT, to reduce uncertainty in the annualized estimate. If the fan or pump unit speed is primarily determined by the facility operating schedule, a measurement period duration of one month or longer tends to be appropriate. If there is no seasonality to the operating schedule; e.g., summer session in schools or peak production month in manufacturing, monitoring potentially can start at any time of the year. If seasonality exists, measurement across multiple seasons may be warranted to capture variation and reduce uncertainty. If long-term VFD speed is available from a control system, short-term kilowatt metering can be obtained concurrently with speed data to develop the relationship between speed and kilowatts. This may allow a shorter metering period when the kilowatts then can be calculated for a longer term using the relationship applied to the speed trend data. If the system has two identical pumps that alternate operation, make sure both are metered, or that only one is allowed to operate during the metering period.

- **Alternative power measurement.** In lieu of true power trending, it is acceptable to use current trends combined with one-time true power measurement of the circuit powering the VFD at three levels of percent speed, including one at 100% speed.

Care should be taken with the acquisition of any power measurements and should conform to Chapter 9: Metering Cross-Cutting Protocols.

- **Trend log of VFD facility OAT.** These data may be obtained from the facility's control system, if it can be programmed to record OAT at 15-minute intervals. Data must be collected for the same period as the VFD current trend log. These data are required only if fan or pump speed is primarily a function of OAT (such as for a heating or cooling units). If OAT data are not available from the facility's control system or appear unreliable, an OAT data recorder should be installed to create this trend log.
- **TMY OAT.** For sites that are OAT dependent, typical hourly OAT data for the weather station nearest or most representative of the M&V site should also be obtained for extrapolating the measurement period savings to a typical operation year.

4.1.3 Savings Estimation Steps

This Performance Curve Method estimates energy savings based on power measurements taken post-VFD-installation, together with fan/pump performance specifications and site operating characteristics. The basis for the calculation is that the post-installation power trend, combined with site system specification, allows the derivation of the post-installation flow trend. This flow trend is assumed to apply in the baseline case (with exceptions for recirculation, which can be added to the baseline flow), where it is used to derive the baseline power trend from the pump performance curve.

The post-operating curve (system curve) is assumed to not vary. All valves either have been removed or are fixed. The system curve is specified with two points: the static head and an operating point based on the equation:

$$h = h_0 + aQ^C \tag{2}$$

Where,

h_0	=	static head, or the head at zero flow
h	=	head
Q	=	flow
Exponent c	=	defaults[7] to 1.7
a	=	correlating coefficient

[7] The default value of 1.7 is based on a consensus of the RTF subcommittee responsible for the technical review of this protocol. The theoretical value based on the affinity laws is 2.0, with actual in situ values less than the theoretical. The actual system curve for a specific system configuration can be determined through flow and pressure measurements, but is beyond the scope of this protocol.

Baseline power is derived from the flow profile. Flow, head, and efficiency points from the performance curve are used to correlate a flow-to-power relationship. To account for cases where a recirculation flow exists in the baseline for a pump system that does not exist in the post-installation period, the minimum flow or constant circulation flow can be used to modify the flow profile.

Savings from the period of measurement are annualized based on annual schedules or a correlation with OAT, if relevant.

The general overall equation describing this method is:

$$\text{Annual Savings} = \sum_{Bins}(Baseline\ kW - Installed\ kW) \times Bin\ Hours \quad (3)$$

Where,

Annual Savings	=	is in kWh
Baseline kW	=	the calculated kW averaged into the appropriate bins (OAT, production, schedule) and extrapolated to the full range of the bin parameter(s) for the site
Installed kW	=	the metered kW averaged into the appropriate bins (OAT, production, schedule) and extrapolated to the full range of the bin parameter(s) for the site
Bin Hours	=	the number of hours in each parameter bin

The specific steps are as follows. Steps 4a and 4b are mutually exclusive alternatives.

Step 1. Derive flow versus power relationship for post-installation period.

System curve. Use the full-flow operating point to correlate an equation for the system curve (flow versus pressure) as the parabola (or lower order equation) from the static head value through the point on the fan or pump curve matching the operating point.

Flow versus power curve. The fan or pump will operate along the system curve as the VFD changes fan or pump speed. Fan/pump efficiency may vary along this curve and is included in the flow versus power calculations. Derive an equation for flow as a function of power (kilowatts) along this curve. Motor efficiency and VFD efficiency are based on default relationships (U.S. Department of Energy [DOE] tables[8]) according to motor percent load.

Step 2. Derive flow versus power relationship for baseline period.

Flow versus power curve. Assuming constant speed, the fan/pump will operate along the fan/pump curve. Using the fan/pump curve data points, derive an equation for power as a function of flow along the fan/pump curve. The manufacturer's curve usually provides power as brake horsepower, so the values will need to be converted to kilowatts and divided by the motor efficiency to obtain comparable kilowatt values.

Step 3. Compute savings for trend log intervals.

Post-installation flow. Calculate flow as a function of kilowatts, using the equation derived from the system curve.

[8] Table from DOE Motor Tip Sheet 11, June 2008.

Baseline flow. The assumption is that the baseline flow profile is identical to the post-installation flow profile, with the exception of recirculation adjustments to the baseline flow.

Baseline kilowatts. Calculate kilowatts as a function of flow, using the equation derived from the fan/pump curve.

Savings for trend log period. Calculate the kilowatt savings profile as the difference between baseline kilowatts and post-installation kilowatts.

Step 4a. Annualize savings: fan/pump speed determined by OAT.

This method for annualizing savings assumes the load can be modeled as driven by OAT.

Average savings by trend log bin. Average kilowatt savings by 2°F temperature bins for all trend log intervals during operating hours, as defined by facility operation schedules. If the facility has more than one operation mode (which determines fan/pump speed), temperature bin averages are separately computed for each operation mode.

Operating hours by TMY bin. Divide the 8,760 TMY OAT data into 2°F temperature bins and compute the frequency of annual operating hours for each bin, as defined by facility operation schedules.

Average savings by TMY bin. TMY average bin savings equal trend log average bin savings for each matching bin. Extrapolate average savings for TMY bins that do not have trend log data. A linear change point correlation model often produces a good relationship between bin temperature and savings. No bin value is allowed to exceed the rated fan/pump motor kilowatts.

Saving by bin. For each TMY bin, multiply the average bin savings by the number of operating hours in each bin to see kilowatt-hour savings in each bin.

Annual savings. Sum the kilowatt-hour values across TMY bins.

Alternatively, the savings can be averaged into 1°F temperature bins, and the savings can be applied to 8,760 hourly TMY temperatures to obtain a complete profile.

Step 4b. Annualize savings: fan/pump speed determined by facility schedule.

This method makes two assumptions: (1) there is a strong correlation between schedule periods and savings; and (2) power trends for the post-installation period are available for all schedule periods.

Average savings for trend period. For the trend log period, average the savings for each operation mode, as determined by facility operation schedule (Section 4.1.2.2).

Annual operating hours. Determine the number of operating hours for each operating mode.

Savings by operating mode. Multiply the number of annual operating hours by the average saving for each operating mode.

Annual savings. Sum savings across operating modes.

4.2 Default Curve Method

This section describes the method for determining the baseline consumption from using the appropriate default curve that describes the flow-versus-power relationship of the fan or pump. This method is consistent with IPMVP Option A (Isolation Retrofit, Key Parameter Measurement).

This relationship for the VFD is assumed to be determined from metering data. However, a default curve is also available if either flow or power cannot be metered for the VFD system.

The primary application of this method is for conditions where the system curve changes due to system damper or valve adjustments downstream of the fan or pump. However, because default curves, [9] instead of curves specific to the fan or pump, are used to define the baseline operation, the method is less accurate. The best method would be to meter baseline power and flow to define the *in situ* performance directly, but this requires extensive and potentially costly baseline measurements and therefore is not generally done. These default curves are industry standard practice, are readily available, and are used in DOE-based hourly simulation models such as eQUEST and EnergyPlus.

4.2.1 Eligible Projects

The advantage of the Default Curve Method is that it is applicable to a much broader range of fan or pump configurations and control schemes than the Performance Curve Method. For each valid combination, the relationship of flow to power is described by a quadratic equation of the form.

$$\text{Flow} = a + b \times (\text{Power}) + c \times (\text{Power})^2 \qquad (4)$$

Where,

Flow = the decimal percent of full flow

Power = the decimal percent of full power

a, b, c = correlation coefficients

Tables 1 and 2 list the correlation coefficients for the applicable fan or pump and control type combinations.

Table 1. Fan Default Curve Correlation Coefficients

Fan Control Strategy	Coeff	Fan Type		
		Forward Curved	Backward Curved or Airfoil	Vane Axial
Discharge dampers	A	0.190667	0.227143	n/a
	B	0.310000	1.178929	n/a
	C	0.500000	–0.410714	n/a
Inlet vane	A	0.339619	0.584345	n/a
	B	–0.848139	-0.579167	n/a
	C	1.495671	0.970238	n/a
Variable pitch	A	n/a	n/a	0.3544
	B	n/a	n/a	–0.9691
	C	n/a	n/a	1.6104
VFD	A	0.219762	0.219762	0.219762
	B	–0.874784	–0.874784	–0.874784
	C	1.652597	1.652597	1.652597

[9] The curves were developed in the early 1970s by Westinghouse (no date), although the raw data and information describing the conditions of the data collection have not been published.

Table 2. Pump Default Curve Correlation Coefficients

Pump Control Strategy	Coefficient	Pump Type Centrifugal
Throttle valve	A	0.55218
	B	0.63701
	C	–0.18996
VFD	A	0.219762
	B	–0.874784
	C	1.652597

Limitations remain for the fan or pump eligibility when the Default Curve Method is used:

- Loads served are similar pre- and post-installation, but airflow or water flow is varied by a different mechanism, as listed in Table 2. If VFD controls are overridden and manually set, the only valid baseline would be if the flow were also manually set in a similar fashion.
- The method is applied to a single fan/pump. Multiple fans/pumps must be treated separately. Fans/pumps configured in parallel that are controlled to operate at the same speed can be evaluated by this method, but fans in series would be excluded. Backup fans/pumps or multiple fans/pumps where the same number operate in parallel can also be evaluated with this method.
- Fan or pump motors must be single-speed motors.
- Baseline control strategies that are not eligible for this method include bypass or cycling.

4.2.2 Data Collection Requirements

4.2.2.1 Fan/Pump, Motor and Variable-Frequency Drive Specifications

- **Fan/pump motor horsepower.** This information is obtained from the motor nameplate.
- **Fan/pump motor revolutions per minute.** This information is obtained from the motor nameplate.
- **Motor enclosure type.** This information is obtained from the motor nameplate.
- **Motor efficiency.** This information is obtained from the motor nameplate.
- **VFD rated efficiency.** This information is obtained from the VFD or from manufacturer's specifications.

4.2.2.2 Fan/Pump Operations

- **Determinants of fan/pump speed.** Possible determinants are: (1) facility operation schedule, (2) OAT combined with operation schedule, and (3) production level.
- **Typical OAT.** Determine if OAT is a significant determinant through discussion with the facility operator and consideration of the types of loads being served by the fan/pump.
- **Facility operation schedule.** Obtained from facility operator. If the fan/pump unit has different operation modes determined by the facility's operation schedule (e.g., setback of

flow during night and weekend hours), determine the period for each mode, defined as needed by hour of day, day of week, and season. This method requires that all schedule modes be metered.

- **Facility production level.** This information is obtained from the facility operator. If the facility's production schedule has established different operation modes for the fan or pump unit (e.g., two production lines for one work shift and one line for other shift), determine the period for each production level, defined as needed by hour of day, day of week, and season. This method requires that all production modes be metered.
- **Weather station.** If OAT is a significant determinant, identify the TMY weather station that is closest to the project site.

4.2.2.3 Post-Period Measurements

- **True RMS power.** This protocol prefers a trend log of true poly-phase RMS power for the circuit powering the VFD, and 15-minute intervals are desired for the trend data. In general terms, a measurement period should be long enough to observe significant variation of dependent operating variables, such as OAT, to reduce uncertainty in the annualized estimate. If the fan or pump unit speed is primarily determined by the facility operating schedule, a measurement period of 1 month or longer tends to be appropriate. If there is no seasonality to the operating schedule; e.g., summer session in schools or peak production month in manufacturing, monitoring potentially can start at any time of the year. If seasonality exists, measurement across multiple seasons may be warranted to capture variation and reduce uncertainty. If long-term VFD speed is available from a control system, short-term kilowatt metering data can be obtained concurrently with speed data to develop the relationship between speed and kilowatts. This may allow a shorter metering period when the kilowatts then can be calculated for a longer term using the relationship applied to the speed trend data. If the system has two identical pumps that alternate operation, make sure both are metered, or that only the one is allowed to operate during the metering period.
- **Alternative power measurement.** In lieu of true power trending, it is acceptable to use trending of current combined with one-time true power measurement of the circuit powering the VFD at three levels of percent speed, including one at 100% speed.
- **VFD speed measurement.** Coincident with the power trending, the VFD percent speed or hertz should also be trended. This may be available from the building control system or as a control point output from the VFD. If it is not possible or reasonable to trend VFD speed, the default curve (Equation 4) will be used to determine an airflow fraction based on the power trend data. If both power and speed trends are available, these will be used as the actual VFD performance curve instead of the default curve.
- **Maximum power point.** This is the power, in kilowatts, at 100% speed for the VFD. The Default Curve Method is based on the decimal fraction of flow and power relative to this maximum point. This value can be determined in one of the following ways.
    - **One-time true RMS power measurement with the system operating at 100% speed.** Operation should be consistent with normal parameters. If the VFD is manually set to 100% speed (overriding normal control), the power value may not

be accurate because operation under these conditions will be at static pressure greater than normally controlled.

If neither power measurement at 100% speed nor speed trending is available, the following methods can be used to determine the power at 100% speed.

- **Take one-time power measurement, record the speed, and convert to fraction of full speed.** Calculate power at 100% speed by using the one-time power and fraction of full speed measurement with the VSD default curve.
- Testing and balancing documents may contain measured power at 100% speed.
- **Design documents or mechanical schedule from as-built plan set will show an anticipated brake horsepower value at full flow, which is the horsepower required for the fan/pump without the motor.** To get the motor horsepower energy demand (motor input power), multiply the brake horsepower by 0.746 to convert to kilowatts and divide by motor efficiency (from nameplate or from MotorMaster Plus based on motor horsepower, rated speed, and enclosure type) to achieve an approximate value for maximum kilowatts.

Care should be taken with the acquisition of any power measurements and should conform to Chapter 9: Metering Cross-Cutting Protocols.

- **Trend log of VFD facility OAT.** These data may be obtained from the facility's control system if it can be programmed to record OAT at 15-minute intervals. Data must be collected for the same period as the VFD current trend log. These data are required only if fan/pump speed is primarily a function of OAT (such as for a heating or cooling units). If OAT data are not available from the facility's control system or appear unreliable, an OAT data recorder should be installed to create this trend log.
- **TMY OAT.** For sites that are OAT-dependent, typical hourly OAT data for the weather station nearest or most representative of the M&V site should also be obtained to extrapolate the measurement period savings to a typical operation year.

4.2.3 Savings Estimation Steps

The Default Curve Method estimates energy savings based on trended measurements taken of the post-VFD-installation.

The general overall equation describing this method is:

$$\text{Annual Savings} = \sum_{Bins}(Baseline\ kW - Installed\ kW) \times Bin\ Hours \quad (5)$$

Where,

Annual Savings	=	in kWh
Baseline kW	=	the calculated kW averaged into the appropriate bins (OAT, production, schedule) and extrapolated to the full range of the bin parameter(s) for the site
Installed kW	=	the metered kW averaged into the appropriate bins and extrapolated to the full range of the bin parameter(s) for the site

Bin Hours = the number of TMY hours in each parameter bin

The specific steps are as follows.

Step 1. Derive flow versus power relationship for post-installation period.

- Convert the trend log data to decimal percent of full flow values by dividing the speed trend values (hertz) by 60 Hz and the power value by the power at 100% speed.
- If speed trends are available, correlate the fractional flow values to the fractional power values to obtain an *in situ* VFD curve.
- If no VFD speed trends are available, calculate the decimal percent flow for each power value using the VFD default curve.

Step 2. Annualize flow fractions determined by OAT.

This method for annualizing savings assumes the load (flow fraction) can be modeled as driven by OAT.

Average flow by trend log bin. Average fractional flow values by 2°F temperature bins for all trend log intervals during operating hours, as defined by facility operation schedules. If the facility has more than one operation mode (which determines fan or pump speed), temperature bin averages are separately computed for each operation mode.

Operating hours by TMY bin. Divide the 8,760 TMY OAT data into 2°F temperature bins and compute the frequency of annual operating hours for each bin, as defined by facility operation schedules.

Average flow fractions by TMY bin. TMY average bin flow fractions equal trend log average bin flow fractions for each matching bin. Extrapolate average flow fractions for TMY bins that do not have trend log data. For higher temperature bins, extrapolation by a linear equation fitted to the trend log bins above 57°F works well; this is also true for lower temperature bins by a linear equation fitted to the bins below 57°F. No bin value is allowed to exceed the full flow value of 1.0.

VFD power by bin. For each TMY bin, calculate the VFD power by using the flow-to-power correlation developed from the trend data. If no speed trend data are available, use the VFD default equation. The trend data will inherently include the part load VFD efficiency, and if the default equation is used the part load efficiency is also included, because it was based on measured data.

Baseline power by bin. For each TMY bin, calculate the baseline power by using the flow with the appropriate default equation; i.e., the equation along with Tables 1 and 2 provide the expression for flow as a function of power. The equation needs to be rearranged and solved for power as a function of flow to calculate baseline power. Alternatively, the default curves can be used to generate flow and power data points and the data can be correlated using a quadratic equation in the form of power as a function of flow for the desired baseline fan or pump type and control type. The resulting power must be multiplied by the rated VFD efficiency to obtain baseline power without the VFD attached. If the rated efficiency is not available, a default efficiency of 97% may be assumed.

Savings by bin. For each TMY bin, calculate the savings as the difference between baseline and VFD power.

Annual savings. Sum the kilowatt-hour values across TMY bins.

Alternatively, 1°F temperature bins can be used instead of 2°F bins, and the savings can be applied to 8,760 hourly TMY temperatures to obtain a complete annual profile.

Step 3. Annualize savings: fan/pump speed determined by facility schedule.

This method makes two assumptions: (1) there is a strong correlation between schedule periods or production level and flow fractions; and (2) trend data for the post-installation period are available for all schedule periods or production levels.

Average flow fractions for trend period. For the trend log period, average the flow fractions for each operation mode, as determined by facility operation schedule or production level.

Annual hour of schedule or production bins. Determine the number of operating hours for each operating mode or production level.

VFD power. For each schedule or production level bin, calculate the VFD power by using the flow-to-power correlation developed from the trend data. If no speed trend data are available, use the VFD default equation.

Baseline power. For each schedule or production level bin, calculate the baseline power by using the flow with the appropriate default equation; i.e., the equation along with Table 1 or 2 that describes the baseline fan/pump type and control type.

Savings by bin. For each schedule or production level bin, calculate the savings as the difference between baseline and VFD power.

Annual savings. Sum the kilowatt-hour values across bins.

4.3 Regression Modeling Direction

To calculate normalized savings, whether following the IPMVP's Option A, Option C, or Option D, the baseline and reporting period regression model must be developed for most projects. [10] This section is for general reference when developing correlations or extrapolation in the sections above. Three types of analysis methods can be used to create a model:

- Linear regression: For one routinely varying significant parameter (e.g., OAT).[11]
- Multivariable linear regression: For more than one routinely varying significant parameter (e.g., OAT, process parameter).
- Advanced regression: Such as polynomial or exponential.[12]

When required, these models should be developed in accordance with best practices, and they should be used only when they are statistically valid (see Section 4.3.1). If no significant

[10] This could either be (1) a single regression model that uses a dummy variable to differentiate the baseline/reporting period data or (2) two independent models for the baseline and reporting period respectively.

[11] One of the most common linear regression models is the three-parameter change point model. For example, a model that represents cooling electricity consumption would have one regression coefficient that describes nonweather-dependent electricity use; a second regression coefficient that describes the rate of increase of electricity use with increasing temperature; and a third parameter that describes the change point temperature, also known as the balance point temperature, where weather-dependent electricity use begins.

[12] Advanced regression methods might be required if a chiller plant is providing cooling for manufacturing or industrial processes.

independent variables are present, no model is required, because the calculated savings will be inherently normalized.

4.3.1 Testing Model Validity

To assess the accuracy of the model, review the parameters listed in Table 3 (EVO 2012).

Table 3. Model Statistical Validity Guide

Parameter Evaluated	Description	Suggested Acceptable Values
Coefficient of determination (R2)	A measure of the extent to which variations in the dependent variable from its mean value are explained by the regression model.	> 0.75
T-statistic	An indication of whether the regression model coefficients are statistically significant.	> 2
Mean bias error	An indication of whether the regression model overstates or understates the actual cooling load.	Will depend on the project, but generally < ± 5%

If any of these parameters fall outside their acceptable range, attempts should be made to enhance the regression model by increasing or shifting the measurement period in one of these three ways: (1) incorporating more data points, (2) including independent variables that were previously unidentified, or (3) eliminating statistically insignificant independent variables.

After enhancement attempts, if the model is still outside the suggested range, this indicates that parameter coefficients are quite poorly determined and that normalized consumption will have relatively high statistical prediction error. Ordinarily such a model should not be used for normalization, unless the analysis includes appropriate statistical treatment of this prediction error. Discussion of how to proceed in such circumstances is outside the scope of these guidelines.

5 Sample Design

Uniform Methods Project Chapter 11: Sample Design describes general sampling procedures that should be consulted if the VFD project population is sufficiently large or if the evaluation budget is constrained.

Ideally, stratified sampling should be undertaken by partitioning VFDs by application (fan, pump, or process versus HVAC load), operating hours, size, and/or the magnitude of claimed (*ex-ante*) project savings. This stratification ensures that sample findings can be extrapolated confidently to the remaining project population.

The confidence and precision-level targets that influence sample size are typically governed by regulatory or program administrator specifications.

6 Other Evaluation Issues

When claiming net program VFD measure impacts, the following evaluation issues should be considered in addition to first-year gross impact findings:

- Net-to-gross estimation
- Realization rates.

6.1 Net-to-Gross Estimation

The cross-cutting chapter, *Estimating Net Savings: Common Practices*, discusses various approaches for determining net program impacts. Best practices include close coordination between gross and net impact results and teams collecting site-specific impact data to ensure that there is no double-counting of adjustments to impacts at a population level.

6.2 Realization Rates

For program-induced projects, realization rates are calculated as the evaluated (*ex-post*) gross savings/claimed (*ex-ante*) gross savings.

References

Cadmus Loadshape (2014). *Variable Speed Drive Loadshape Project – Draft Report.* Submitted to Northeast Energy Efficiency Partnerships August 2014.

EVO (2012). *International Performance Measurement and Verification Protocol – Concepts and Options for Determining Energy and Water Savings Volume 1*. January 2012. Washington, D.C.: Efficiency Valuation Organization.

SBW Pump (2012). *Standard Savings Estimation Protocol for Pump VFD.* Submitted to the Regional Technical Forum October 17, 2012. http://rtf.nwcouncil.org/subcommittees/fanvfd/. San Francisco, CA: SBW Consulting, Inc.

SBW Fan (2012). *Standard Savings Estimation Protocol for Fan VFD.* Submitted to the Regional Technical Forum November 29, 2012. Accessed October 30, 2014: http://rtf.nwcouncil.org/subcommittees/fanvfd/. San Francisco, CA: SBW Consulting, Inc.

Westinghouse (no date). Bulletin B-851, F/86/Rev-CMS 8121. *Flow Control*.

Resources

Bernier, M.A.; Bourret, B. (1999). "Pumping Energy and Variable Frequency Drives." *ASHRAE Journal*. December 1999, p. 37.

California Energy Commission (2007). *Nonresidential Alternative Calculation Method for the 2008 Building Energy Efficiency Standards for Residential and Nonresidential Buildings.* November 2007.

Chapter 19: HVAC Controls (DDC/EMS/BAS) Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in August 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in November 2014.

Jeff Romberger
SBW Consulting, Inc.
Bellevue, Washington

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68575
September 2017

Contract No. DE-AC36-08GO28308

Chapter 19: HVAC Controls (DDC/EMS/BAS) Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in August 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in November 2014.

Jeff Romberger
SBW Consulting, Inc.
Bellevue, Washington

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68575
September 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter author wishes to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Dan Barbieri of DNV-GL
- David Jacobson of Jacobson Energy Research, LLC.

Suggested Citation

Romberger, Jeff. (2017). *Chapter 19: HVAC Controls (DDC/EMS/BAS) Evaluation Protocol*, *The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68575. http://www.nrel.gov/docs/fy17osti/68575.pdf

Acronyms

BAS	building automation system
CDD	cooling degree day
DDC	direct digital controls
EMS	energy management system
HDD	heating degree day
HVAC	heating, ventilating, and air conditioning
IPMVP	International Performance Measurement and Verification Protocol
M&V	measurement and verification
OAT	outside air temperature
RMS	root mean square
TMY	typical Meteorological Year

Table of Contents

List of Tables

1 Measure Description

The HVAC Controls Evaluation Protocol is designed to address evaluation issues for direct digital controls/energy management systems/building automation systems (DDC/EMS/BAS) that are installed to control heating, ventilation, and air-conditioning (HVAC) equipment in commercial and institutional buildings. (This chapter refers to the DDC/EMS/BAS measure as HVAC controls.) This protocol may also be applicable to industrial facilities such as clean rooms and labs, which have either significant HVAC equipment or spaces requiring special environmental conditions.[1]

This protocol addresses only HVAC-related equipment and the energy savings estimation methods associated with installing such control systems as an energy efficiency measure. The affected equipment includes:

- Air-side equipment (air handlers, direct expansion systems, furnaces, other heating- and cooling-related devices, terminal air distribution equipment, and fans)
- Central plant equipment (chillers, cooling towers, boilers, and pumps).

These controls may also operate or affect other end uses, such as lighting, domestic hot water, irrigation systems, and life safety systems such as fire alarms and other security systems.

Considerable nonenergy benefits, such as maintenance scheduling, system component troubleshooting, equipment failure alarms, and increased equipment lifetime, may also be associated with these systems. When connected to building utility meters, these systems can also be valuable demand-limiting control tools. However, this protocol does not evaluate any of these additional capabilities and benefits.

[1] As discussed under the section *Considering Resource Constraints* of the "Introduction" chapter to this UMP report, small utilities (as defined under the U.S. Small Business Administration (SBA) regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

2 Application Conditions of Protocol

The type of HVAC control system to which this protocol applies is common in newly constructed commercial and institutional buildings that are larger than 100,000 ft^2 (PNNL, 2012). However, numerous older buildings have either minimal HVAC controls or older systems with less-efficient control sequences that can benefit from this measure. Many older BASs use pneumatic controls, which are often in disrepair. There is also a significant opportunity for more advanced control systems in smaller buildings. Energy efficiency programs encourage the installation of HVAC controls as retrofits in existing facilities and, in some cases, encourage installation in new construction. Generally, energy codes do not require that DDC/EMS/BAS-type controls be installed; however, energy codes tend to specify minimum HVAC control features, such as time-of-use on/off scheduling and economizer controls on air handlers. Some codes specify significantly more control requirements, such as reset schedules on supply air temperature in air handlers. In instances where code minimum requirements apply to new construction, or to new HVAC systems in major renovation projects, code-required controls should be considered baseline.

Two common program delivery mechanisms are in use around the country.

- **Prescriptive.** This approach usually entails an incentive that is based on an easy-to-calculate building metric (such as the building floor area affected by the HVAC controls) or on the number of qualifying control points. The incentives may vary, based on the type of building where the equipment is installed, because the achieved energy savings tend to be specific to a building's use (e.g., hospitals versus schools).
- **Custom.** This approach also provides an incentive for the HVAC controls that is based on the expected annual energy savings (kilowatt-hours) estimated using a custom calculation tool. The custom calculation approach is often used for facilities that apply incentives to multiple measures in a building. In this circumstance, estimation may be the result of a complete hourly building energy simulation model developed using a program such as eQUEST or EnergyPlus. However, other calculation approaches may be used, such as developing a bin model for the HVAC systems in a building. Custom programs may require measurement and verification (M&V) after the controls are installed to calculate savings and determine incentive amounts based on actual equipment performance.

3 Savings Calculations

This section presents a high-level equation that applies to all HVAC controls measures for calculating gross energy savings.[2] Detailed direction on how to apply this equation is presented in the Measurement and Verification Plan section of this protocol.

Energy savings are determined using the following general equation (EVO 2012):

Equation 1

$$\textit{Energy Savings} = (\textit{Baseline Energy} - \textit{Reporting Period Energy}) \pm \textit{Routine Adjustments} \pm \textit{Non-Routine Adjustments}$$

Where,

Energy Savings	= First-year energy consumption savings.
Baseline Energy	= Pre-implementation consumption.
Reporting Period Energy	= Post-implementation consumption.
Routine Adjustments	= Adjustments made to account for routinely changing independent variables (variables that drive energy consumption). Savings should be normalized to typical meteorological year (TMY) weather data as well as other significant independent variables (e.g., occupancy, production data), if applicable. If first-year energy savings are desired, savings should be normalized to the actual weather for the 12-month period following commissioning of the new controls.
Non-Routine Adjustments	= Adjustments made to account for parameters that typically are not expected to change during the implementation period. If these parameters change *and* this change influences the reporting period energy use, they should be accounted for (e.g., changes to a facility's building envelope during implementation of an HVAC controls measure). This would have to be considered if savings were verified using Options C or D of the International Performance Measurement and Verification Protocol (IPMVP) (EVO 2012) or calibrated simulation per ASHRAE Guideline 14-2002 (ASHRAE 2002).[3]

HVAC controls demand savings is not a straightforward extension of verified consumption savings (unlike lighting retrofits, for example, where established load savings profiles can easily be applied to consumption savings data). For HVAC controls projects, the load savings profiles vary, depending on the distribution of the measure types implemented and are weather dependent. These complications are accounted for in the M&V methods described in Section 4.

---

[2] This protocol focuses on gross energy savings and does not include other parameter assessments, such as net-to-gross, peak coincidence factors, or cost effectiveness.

[3] Whole-facility consumption analysis

4 Measurement and Verification Plan

This section contains four approaches for determining the energy savings resulting from the HVAC controls measure, and provides guidance on how and when to use each.

Two methods use pre- and post-installation data and the other two use post-installation data only.

- The first method (End Use Regression Model) is more accurate than the second, using pre- and post-installation metered data of the affected end uses. However, it has limitations caused by metering requirements and the necessity for fewer complication factors. The method is appropriate only for retrofit of HVAC controls.
- The second method (Consumption Data Analysis) is similar to the first, but it is much simpler and cheaper to conduct. However, because this method uses whole building consumption data (billing records, whole building interval or AMI data), it is also less accurate and typically requires that expected savings are greater than 10% of base year energy use to be separated from the noise. The method is appropriate only for retrofit of HVAC controls.
- The third method (Bin Model Calculations) is useful when pre-installation metering is not possible. This method can be used for most situations; however, it can be expensive to conduct for large, complicated systems unless a model is available from the ex-ante analysis or the evaluator has a model available that can be easily adapted. The method can be used for both retrofit of HVAC controls or for new construction.
- The fourth method (Calibrated Simulation) is appropriate for complex facilities, and it can be reasonably cost effective if the building simulation model is available from the ex-ante savings estimate documentation. The method can be used for both retrofit of HVAC controls or for new construction.

The End Use Regression Model is the most accurate method and is recommended for M&V of HVAC controls measures. However, it is rarely used because it has extensive pre- and post-retrofit metering requirements and potential complications from interactions of other measures concurrently installed. The premetering requirement can delay the installation and add significant cost to the M&V process. The other methods are all commonly used and M&V method selection is often based on the available calculation algorithm used to determine the ex-ante energy savings. Ultimately, the choice of method depends on the cost of performing the measurements, the magnitude of energy savings, and the scope and complexity of the measure.

System Commissioning. *Generally programs require commissioning of control systems prior to incentive payment. It is very important that the installed controls have been commissioned and are functioning as intended before M&V activities commence. It is also important that the M&V activities do not influence the customer or utility program behavior prior to M&V, since the objective is to provide an independent evaluation of measure performance. M&V should commence after sufficient time has passed to ensure full opportunity for commissioning but before the occurrence of supplemental control sequence changes not originally specified.*

4.1 Baseline Definition Considerations

For the measure evaluation to be consistent with program requirements, it is important to define clearly the baseline conditions. The two primary areas to consider when defining the baseline conditions are program requirements and multiple measure installation (discussed in Section 4.1.1 and Section 4.12, respectively). These considerations also impact the selection of the savings calculation method.

4.1.1 Program Requirements

The program under which the HVAC controls are incented often has specific rules about a measure's eligibility. For custom programs, the incentive payment is based on the estimated energy savings. Also, these programs often have specific requirements for the baseline definition as it relates to estimating the claimed savings. Some custom programs base the final incentive amount on the actual energy savings after a measure is installed, and those savings are often determined by required M&V process and a recalculation of savings.

Common eligibility criteria for new construction specify only that the HVAC control features exceed energy code minimum requirements. Therefore, the prevailing energy code must be examined carefully before a list of eligible controls can be developed for a project. For example, with retrofit applications, savings are often based on the pre-installation control of the affected systems. So it is important to determine whether the energy code was—or should have been—triggered by the retrofit, as this might impact the baseline estimate as if the project were new construction. Furthermore, the code to use for the baseline can be difficult to determine for projects with long timelines between design and construction. For example, some states apply the code in effect at the time of 100% construction drawings; others use the code in effect during plan submission at 60% drawings. This can make a significant difference in savings estimates for new construction or major renovations.

Also, some program rules specify that broken controls (or controls that are in place but are overridden in the pre-retrofit period) are not eligible and should not be considered in the savings estimate. These types of retrocommissioning issues include:

- An economizer that has dampers stuck in one position due to a failed damper motor
- A time clock for on/off scheduling that is not programmed (or has had all the "off" pegs removed) and allows the system to run all the time.

Although these types of retrocommissioning issues will likely be allowed under a retrocommission program, if they were completed under a custom HVAC program, the program rules may not allow them.

4.1.2 Multiple Measure Installation

For a major renovation, the HVAC controls measure is often only one of several in an overall package of measures. The package may include replacing constant-volume with variable-volume air handlers and replacing a chiller plant. In that instance, significant interactions between the measures need to be considered if the evaluation encompasses the HVAC controls measure only ***or*** if savings must be evaluated individually for each measure.

Although this protocol does not address the interactions of measures, it contains recommendations for the appropriate evaluation method to account for interactions. The first method, the End Use Regression Model, is discussed in detail with step-by-step descriptions. The other three methods are discussed in general terms only, because they are less conducive to being described in terms of a uniform method.

4.2 End Use Regression Model Method

Consistent with IPMVP Option B (Isolation Retrofit, All Parameter Measurement), this method uses measured pre- and post-installation metering of kilowatt consumption of all affected end uses (heating, cooling, fans, pumps, other auxiliary). The metered data are averaged into temperature bins that are based on the outside air temperature (OAT) (obtained from concurrent metering). The model is then adjusted for weather differences by applying TMY weather data to the measured data and extrapolating to all temperatures.

A significant advantage of this method is that the analyst does not need to know how to describe the control features, either in an engineering equation (as required for the Bin Model Calculation Method), or in a simulation model (as required for the Calibrated Simulation Method). Some control features are difficult to express with these methodologies.

The general overall equation describing this method is:

$$Annual\ Savings = \sum_{End\ Uses} \sum_{Temp\ Bins} (Baseline\ kW - Installed\ kW)\ \times Bin\ Hours \quad (2)$$

Where,

Annual Savings is in kWh,

Baseline kW is the metered kW averaged into temperature bins and extrapolated to the full range of TMY weather for the site,

Installed kW is the metered kW averaged into temperature bins and extrapolated to the full range of TMY weather for the site,

Bin Hours are the number of TMY hours in each temperature bin.

The specific calculation steps are as follows.

Step 1. Define the system boundary

In defining the boundary around the equipment in the evaluation, include all the equipment directly impacted by the installed HVAC controls. An example of direct impact is the addition of demand ventilation controls to an air handler.

Also, include indirectly impacted equipment if this inclusion is expected to have a greater than 5% effect on the total savings. Examples of indirectly affected equipment are the chiller and boiler serving the air handler with the demand ventilation control if there are resulting changes in heating and cooling loads.

It may be appropriate to include the boiler but not the chiller when a building is located in a cold climate (where cooling energy is a very small percentage of total building consumption and heating energy is a much larger percentage). In a hot climate, the opposite would be true.

Step 2. Collect the data

Collect these data for the evaluation.

- **HVAC load determinants.** In most cases, the heating and cooling loads will be a function of OAT. Identify the TMY weather station that is closest to the project site. The weather data are needed to normalize both pre- and post-energy consumption and thus eliminate weather year differences.
- **Facility operations schedule.** Determine the period for each mode (defined, as needed, by hour of day, day of week, and season), because this method requires that metered data be collected during all schedule modes. If the HVAC systems have different operation modes, determine by the facility's operation schedule (e.g., setback of space temperature set point during night and weekend hours).
- **Equipment inventory.** Obtain nameplate information for each control system's affected equipment within the system boundary.

Step 3. Perform metering

Meter equipment to obtain the following information.

- **True root mean square (RMS) power**. For this protocol, it is preferable to have a trend log (noting the data in 15-minute intervals) of true poly-phase RMS power for all circuits powering the desired end uses. If the system load is primarily determined by OAT, the measurement period must be sufficient for capturing the system's operation during a range of outside temperatures. The metering periods must also span seasonal changes, if any, in the operating schedule. Some HVAC control systems have a power-trending function for some equipment. If using this function, take a one-time power measurement to verify the accuracy of the control system values. If these values are off, develop a calibration curve to adjust the values.
- **Alternative power measurement**. In lieu of true power trending, it is acceptable to trend the electrical current combined with a one-time true power measurement at three load levels within the typical operating range of the equipment.
- **OAT.** Trend the OAT concurrently with the power measurements. This information is likely available from the control system; however, check the values for accuracy. Alternatively, deploy a temperature logger to trend OAT.

When acquiring power measurements, take care that the effort conforms to the metering cross-cutting protocols in Chapter 9.

Step 4. Calculate the savings

Complete the following activities separately for the pre- and post-installation metering data. If more than one metering channel is recorded for each end use, sum all the metering data for each end use to create a single trend of values.

Also, to obtain a complete annual profile, 1°F temperature bins can be used instead of 2°F bins, and the savings can be applied to 8,760 hourly TMY temperatures.

- **Average kilowatts by trend log bin.** For each end use, average thekilowatt values by 2°F temperature bins for all trend log intervals during operating hours, as defined by facility operations schedules. If the facility has more than one operation mode, calculate the temperature-bin averages separately for each operation mode.
- **Operating hours by TMY bin.** Divide the 8,760 TMY OAT data into 2°F bins and compute the frequency of annual operating hours for each bin, as defined by facility operations schedules.
- **Average kilowatts by TMY bin.** The TMY average-bin kilowatts equal the trend log average-bin kilowatts for each matching bin. Extrapolate the average kilowatts for TMY bins that do not have trend log data. Plot the kilowatt value versus bin temperature data and then determine the regression equation that best fits through the data that extrapolates to the highest and lowest TMY temperature bin. No bin kilowatt value is allowed to exceed the full equipment kilowatt capacity.
- **Savings by bin.** For each end use and for each TMY bin, calculate the savings as the difference between the baseline estimate and the installed kilowatt values multiplied by the number of hours in the temperature bin.
- **Annual savings.** Sum the kilowatt-hour values across the TMY bins for each end use and then sum the end-use savings into an annual value.

4.3 Consumption Data Analysis Method

Whole-building consumption data analysis is consistent with IPMVP Option C (Whole Building). This option is appropriate when conditions are similar to those of the End Use Regression Model method ***but*** pre-installation end-use metering is not possible or practical. Although this method is much less costly to perform, it is also less accurate. For this method, the HVAC controls measure savings must be large compared to the random or unexplained energy variations at the whole-facility level.[4] Thus, this analysis cannot be undertaken until after a full season or full year of reporting-period billing data is collected.

A billing analysis requires that a full year of billing data be available for the pre- and post-installation years. The monthly energy use of the facility can be correlated to the weather data corresponding to each billing period. Commonly used forms of weather data for correlation include cooling degree days (CDDs), heating degree days (HDDs), and average temperature.

[4] Typically savings should exceed 10% of the baseline energy consumption for a particular meter (e.g., electricity meter) to confidently discriminate the savings from the baseline data when the reporting period is shorter than 2 years (EVO 2012).

Usually temperature will provide the best correlations, but CDD and HDD can also provide good correlations if the best correlation is searched from multiple sets of CDD and HDD values, produced from different base temperatures. This provides weather-specific correlations for the pre- and post-installation periods. The correlations are applied to TMY weather corresponding to each billing period to determine long-term average energy use for the pre- and post-installation periods. Savings are then calculated as the difference between pre- and post-installation annual energy use. If first-year savings are desired, the pre-installation period correlation can be used with the post-installation year weather to obtain energy use corresponding with post-installation year weather conditions.

If interval billing data or AMI data (hourly) is available, the correlations can be greatly improved by using the corresponding hourly weather in the correlations instead of monthly average values. The data can also then be divided into time categories, such as for occupied and unoccupied hours and/or weekday and weekend day types. The correlations can be done for each category. Nonlinear correlations or change point linear correlations may be appropriate to define the models across the entire year.

Note that this method cannot be used for new construction or a major renovation because the baseline whole-building consumption would not be representative of a building constructed to the prevailing energy code. (That is, for a major renovation, the entire building would have to be in compliance with the prevailing energy code.)

4.4 Bin Model Calculations Method[5]

Consistent with IPMVP Option A (Isolation Retrofit, Key Parameter Measurement), this method uses metered key variables of the affected equipment to inform the development of an engineering model that describes system operation. The model is then used to calculate energy consumption for the installed HVAC control system. The baseline consumption is determined by making changes to the model that reflect the baseline system operation.

This method can be used when pre-installation metering data are not available or when there are other significant non-measure changes to the building during either the pre- or post-installation metering period. The types of nonmeasure changes include a significantly different occupancy level, the installation of other conservation measures, or a determination that the baseline is different from the actual pre-equipment operation.

The system boundary is defined through activities similar to those described for Step 1 of the End Use Regression Model method. Also, the data collection effort would encompass both a complete inventory of the equipment within the boundary and the operating sequences of that equipment. (The as-built plans and control system can be very useful for collecting these data.)

Trend data from the control system or from evaluator installed metering equipment should be used to inform the bin model and to calibrate the bin model. Examples of the use of trend data include:

[5] This also called the Inverse Bin method.

- Supply air temperature trends in air handlers to verify temperature reset schedules, such as reset defined by OAT (this requires that OAT be trended also). This can be used in the bin model to define how the supply air temperature varies.
- Fan speed or kilowatt trends to determine the airflow at differing conditions in the model. This might also be a function of OAT. Fan kilowatts can also be used to determine if the model is responding well by comparing the metered fan energy to the bin model calculated fan energy (calibration of the model).

The baseline model will be developed by using the calibrated post-installation model and changing the appropriate control function to baseline conditions. An example is an HVAC controls system change from an air handler with constant supply air temperature control to one that resets supply air temperature based on OAT. The post-installation model should be calibrated with the air handler supply air temperature that varies with OAT as indicated by the trend data. The baseline version of the model will then result from changing the supply air temperature to the constant baseline value.

Because no pre-installation trend data are likely to be available, the baseline conditions need to be verified during the evaluation. This can be done through discussions with staff who are familiar with the baseline system. This may include building maintenance staff or an HVAC maintenance contractor. Baseline control and operation may also be determined by reviewing the original (or most recent pre-installation) mechanical drawings, control sequences, commissioning reports, or balancing reports.

Viewing this method as a uniform method of analysis can be difficult because achieving results that are consistent between analysts is a challenge. This method requires that a site-specific model be developed for the measure; however, engineers are likely to develop the bin model differently, and they may use different trend data to inform the model.

This method is commonly used and can be implemented at a reasonable cost for evaluating difficult or complex situations (such as a heating hot water valve that is leaking and increasing the cooling load), especially when the model has already been developed for the ex-ante savings calculations or when the evaluator has an existing bin model that can be easily adapted. However, if a model is not available, model development cost should be a consideration when choosing an evaluation method.

4.5 Calibrated Simulation Method

Consistent with IPMVP Option D (Calibrated Simulation), this is a good method to use for large, complex facilities because it can handle many different control sequences. It is also a useful approach for modeling multiple measures and accounting for the interactive effects between them.

This method may be cost effective when a model developed for the claimed savings analysis is available to the evaluator. However, it is important to confirm that the model is representative of the actual installed systems. (Unless the model was used for M&V after the installation, it may be different from what was originally anticipated during the claimed savings analysis.)

An essential component of an effective savings estimate is analytical credibility and the ability for a third party to validate it. For this reason, evaluators often exclude proprietary building simulation models, models that prevent the reviewer from seeing the inputs, and models for which the simulation methodology is not published or documented publicly. The U.S. Department of Energy website has an extensive list of whole-building analysis energy simulation tools.[6]

Ideally, the model represents the post-installation conditions and is calibrated to monthly bills with actual weather coincident with the bills. The HVAC control features should then be changed to be consistent with the baseline control features before the model is rerun. The difference between the two runs will be the first-year savings. If long-term typical annual savings estimates are desired, use TMY weather to run the baseline and as-built models.

Issues around obtaining baseline information are the same as those discussed in Section 4.4. It is also important to use trend data from the control system or from evaluator-installed metering equipment to inform and assist in calibrating the model. Space temperature trends can be obtained to confirm the space temperature set points that should be specified in the model. Air handler supply fan kilowatt trends can be used to calibrate the fan energy use that the model predicts. This can greatly enhance the model calibration process by providing some end-use calibration to the model and increase confidence in the overall model calibration. Other data from observation of the control system should also be incorporated into the calibrated model. This will include items such as chilled water temperature set point, air handler operating schedules, and minimum outside air damper settings. Some end uses, such as lighting and plug loads, are more difficult to specify and in some cases may be defined by default characteristics (ASHRAE 2000).

4.6 Other Modeling Considerations

Regression models may be very simple or complex, depending on the significance of the independent variables used. Some general modeling information is included in Section 4.6.1 through Section 4.6.3.

4.6.1 Regression Modeling Direction

To calculate normalized savings—whether following the IPMVP's Option A, Option C, or Option D—develop the baseline and reporting period regression model[7] for most projects (ASHRAE 2004; BPA 2011). The three types of analysis methods used to create a model follow:

- **Linear regression.** For one routinely varying significant parameter (e.g., OAT).[8]

---

[6] http://apps1.eere.energy.gov/buildings/tools_directory/subjects_sub.cfm

[7] This could either be a single-regression model that uses a dummy variable to differentiate the baseline/reporting period data, or two independent models—one for the baseline periods and one for the reporting period.

[8] One of the most common linear regression models is the three-parameter change-point model. For example, a model that represents cooling electricity consumption would have one regression coefficient that describes non-weather-dependent electricity use; a second regression coefficient that describes the rate of increase of electricity use with increasing temperature; and a third parameter that describes the change-point temperature (also known as the balance-point temperature), where weather-dependent electricity use begins.

- **Multivariable linear regression.** For more than one routinely varying significant parameter (e.g., OAT and a process parameter).
- **Advanced regression.** Polynomial or exponential, for example.[9]

When these models are required, develop them in accordance with best practices. Also, use these models only when they are statistically valid (see Section 4.6.2). When there are no significant independent variables, no model is required, because the calculated savings will be inherently normalized.

4.6.2 Testing Regression Model Validity

To assess the accuracy of the model, review the parameters listed in Table 1 (EVO 2012).

Table 1. Model Statistical Validity Guide

Parameter Evaluated	Description	Suggested Acceptable Values
Coefficient of determination (R^2)	A measure of the extent to which the regression model explains the variations in the dependent variable from its mean value.	> 0.75
T-statistic	An indication of whether the regression model coefficients are statistically significant.	> 2
Mean bias error	An indication of whether the regression model overstates or understates the actual cooling load.	< ± 5% (This value is typical, but depends on the project.)

If any of these parameters fall outside of the acceptable range, when possible, attempt to enhance the regression model by:

- Increasing or shifting the measurement period.
- Incorporating more data points. This may include additional metering or obtaining trends from the EMS.
- Including independent variables that were previously unidentified. This may include additional metering or obtaining additional trends from the EMS. It may also be as simple as dividing the available trend data into different categories such as separating Saturdays out because they operate the same hours as weekdays, but at only half the weekday staffing levels.
- Eliminating statistically insignificant independent variables.

After enhancement attempts, if the model is still outside the suggested range, this indicates that parameter coefficients are quite poorly determined and that normalized consumption will have a relatively high statistical prediction error. Ordinarily such a model should not be used for normalization, unless the analysis includes appropriate statistical treatment of this prediction

[9] Advanced regression methods might be required if a chiller plant is providing cooling for manufacturing or industrial processes.

error. Discussion of how to proceed in such circumstances is outside the scope of these guidelines.

4.6.3 Model Calibration

In estimating energy usage for systems and equipment, engineering models rely on thermodynamic, heat transfer, and other physical principles. When it is practical to do so, measure the energy use of the modeled system during the post-installation period. Then compare the estimated energy use (as derived from the model) to the measured use.

To calibrate the model to the measured use, adjust the model inputs or specification, as needed. The objective for this calibration process is to achieve a match between the modeled use and measured use that is within the limits defined by the IPMVP Option D protocol (summarized in Table 2). By applying the model to hourly data and comparing monthly and hourly values of metered data, bin models and statistical models can also be specified to achieve these limits, as determined by ASHRAE Guideline 14-2002 (ASHRAE 2002).

Table 2. Model Calibration Criteria

Data Interval	RMS Error	Maximum Mean Bias Error
Monthly	± 15%	± 5%
Hourly	± 30%	± 10%

5 Sample Design

Consult the UMP *Chapter 11: Sample Design Cross-Cutting Protocol* for a description of general sampling procedures. Use this information when either the HVAC controls measure includes a sufficiently large population of air handlers or the evaluation budget is constrained.

Ideally, use stratified sampling to partition the air handlers by size, type, and operating schedule. This ensures that sample findings can be extrapolated confidently to the remaining project population. The confidence- and precision-level targets that influence sample size are typically governed by regulatory or program administrator specifications.

6 Other Evaluation Issues

When claiming net program HVAC controls measure impacts, consider the following evaluation issues in addition to considering the first-year gross impact findings:

- Net-to-gross estimation
- Realization rates.

6.1 Net-to-Gross Estimation

The cross-cutting chapter, *Estimating Net Savings: Common Practices*, discusses various approaches for determining net program impacts. To ensure adjustments to impacts are not double-counted at a population level, follow the best practices that include close coordination between: (1) staff estimating gross and net impact results, and (2) the teams collecting site-specific impact data.

6.2 Realization Rates

For program-induced projects, divide the claimed (*ex-ante*) gross savings by the evaluated (*ex-post*) gross savings to calculate the realization rates.

7 References

ASHRAE (2002). *Guideline 14-2002 – Measurement of Energy and Demand Savings.* (An update to this guideline is expected in 2014). Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

ASHRAE (2000). RP-1093 – Compilation of Diversity Factors and Schedules for Energy and Cooling Load Calculations. Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

ASHRAE (2004). RP-1050 – Development of a Toolkit for Calculating Linear, Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

BPA. (2011). *Regression for M&V: Reference Guide*. Portland, OR: Bonneville Power Administration.

EVO (2012). *International Performance Measurement and Verification Protocol – Concepts and Options for Determining Energy and Water Savings Volume 1.* Washington, D.C.: Efficiency Valuation Organization.

PNNL (2012). *Energy Savings Modeling of Standard Commercial Building Re-tuning Measures: Large Office Buildings.* Richland, WA: Pacific Northwest National Laboratory.

8 Resources

ASHRAE (1983). *Simplified Energy Analysis Using the Modified Bin Method.* Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Itron, Inc.; KEMA, Inc. (2012). *Evaluation of the 2010-2012 Custom Impact Energy Efficiency Programs for the California IOUs – Procedures for Site-Specific Impact Analysis.*

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/

North America's Home and Building Automation Association 2002. *Best-Practices Guide for Evaluating Intelligent Building Technologies.* IS 2002-28.

Chapter 20: Data Center IT Efficiency Measures Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in October 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in January 2015.

Robert Huang
The Cadmus Group
Portland, Oregon

Eric Masanet
Northwestern University
Evanston, Illinois

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68576
October 2017

Contract No. DE-AC36-08GO28308

Chapter 20: Data Center IT Efficiency Measures Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in October 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in January 2015.

Robert Huang
The Cadmus Group
Portland, Oregon

Eric Masanet
Northwestern University
Evanston, Illinois

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68576
October 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter author wishes to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Dale Sartor of Lawrence Berkeley National Laboratory
- Thomas Bolioli of Terra Novum
- Dan Barbieri of DNV-GL
- Naomis Cole of PECI,
- John Clinger of ICF
- Eric Winkler of ISO New England.

Suggested Citation

Huang, R.; Masanet, E. (2017). *Chapter 20: Data Center IT Efficiency Measures Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68576. http://www.nrel.gov/docs/fy17osti/68576.pdf

Acronyms

EM	efficiency metric
EUL	expected useful life
HDD	hard disk drive
HVAC	heating, ventilating, and air conditioning
IT	information technology
M&V	measurement and verification
MAID	massive array of idle disks
PDU	power distribution unit
PUE	power usage effectiveness
RAID	redundant array of independent disks
SAS	serial attached small computer system interface
SATA	serial advanced technology attachment
SERT	Server Efficiency Rating Tool
SSD	solid-state drive
UPS	uninterruptible power supply
VSD	variable-speed drive

Table of Contents

List of Figures

List of Tables

1 Measure Description

Data centers use about 2% of the electricity in the United States (Koomey 2011); a typical data center has 100 to 200 times the energy use intensity of a commercial building. Data centers present tremendous opportunities -- energy use can be reduced as much as 80% between inefficient and efficient data centers (DOE 2011). Data center efficiency measures generally fall into the following categories:

- Power infrastructure (e.g., more efficient uninterruptible power supplies [UPS], power distribution units [PDUs])
- Cooling (e.g., free cooling, variable-speed drives [VSDs], temperature and humidity set points)
- Airflow management (e.g., hot aisle/cold aisle, containment, grommets)
- Information technology (IT) efficiency (e.g., server virtualization, efficient servers, efficient data storage).

This chapter focuses on *IT measures in the data center* and examines the techniques and analysis methods used to verify savings that result from improving the efficiency of two specific pieces of IT equipment: servers and data storage. The discussion examines options in two categories:

- Using more efficient server and data storage equipment
- Managing servers and data storage equipment to work more efficiently.

Section 1.1 describes some common IT measures that save energy in data centers.

1.1 Server Virtualization

In the past, data center operators ran a single application on each server. This "one workload, one box" approach meant servers ran at a low "utilization rate": the fraction of total computing resources engaged in useful work (EPA undated a). A 2012 *New York Times* article cited two sources that estimated average server utilization rate of 6% to 12% (EPA undated b). Another study stated that the "one workload, one box" approach resulted in 90% of all x86 servers running at less than 10% utilization, with a typical server running at less than 5% utilization (EPA undated b).

Administrators can use server virtualization to run multiple applications on one physical host server, thus consolidating server resources. In other words, multiple virtual servers can work simultaneously on a single physical host server. Therefore, instead of operating many servers at low utilization rates, virtualization combines the processing power onto fewer servers, operating at higher total utilization rates.

1.2 More Efficient Servers

ENERGY STAR®-certified servers have been available since 2009. The ENERGY STAR server specification covers four server form factors (blade, multi-node, rack-mounted, and pedestal) and allows a maximum of four process sockets per server (or per blade or node). ENERGY STAR servers must have the following features:

- Efficient power supplies to limit power conversion losses
- Improved power quality
- Idle power draw limits for rack-mounted or pedestal servers with one or two processors;
- Results of the Server Efficiency Rating Tool (SERT) tests to accommodate comparisons of server efficiency under various usage scenarios
- Ability to measure real-time power use, processor utilization, and air inlet temperatures
- Advanced power management features and efficient components that save energy across various operating states (including idle)
- A Power and Performance data sheet for purchasers; this standardizes key information on energy performance, features, and other capabilities.

On average, ENERGY STAR servers operate about 30% more energy efficiently than standard servers. The servers operate particularly efficiently at low loads because processor power management requirements reduce power consumption when the servers are idle (EPA undated b).

1.3 Data Storage Management

Data storage resource management tools (Clark and Yoder 2008) help data storage administrators more efficiently and effectively provision and manage data storage. This entails using tools to create "maps" and "pools" of available storage across servers and disks, and using these disparate "chunks" of storage as if they operated as one system. These tools include:

- **Automated storage provisioning.** This improves storage efficiency through right-sizing, identifies and reallocates unused storage, and increases server capacity by improving existing storage use (Netapp 2014).
- **Deduplication software.** This condenses the data stored at many organizations by more than 95% by finding and eliminating unnecessary copies. Redundant copies consume more than half the total volume of a typical company's data.
- **Thin provisioning.** This allocates just enough storage just in time by centrally controlling capacity and allocating space only as applications require it. Thus, administrators power only the storage currently in use.
- **Redundant array of independent disks (RAID).** This level is a storage technology that combines multiple disk drive components into a single logical unit. RAID 1 creates a duplicate copy of disk data and doubles the storage and power consumption. For storage that is not mission critical, RAID 5 guards against a single disk drive failure in a RAID set by reconstructing the failed disk information from distributed information on the remaining drives. Requiring only one extra, redundant disk, RAID 5 saves energy, although it sacrifices some reliability and performance. For a 10-disk array, increasing to an 11-disk RAID 5 level (one extra disk) from a 20-disk RAID 1 level (duplicate copy) configuration would save 45% of data storage energy use.

- **Tiering storage.** This automatically stores low-priority data (rarely accessed information) on higher-latency equipment that uses less energy.

1.4 More Efficient Data Storage Equipment

A number of data storage equipment types use less energy (Yoder 2012), including the following:

- **Lower speed drives.** Higher-spin speeds on high-performance hard disk drives (HDDs) (e.g., 15 K rpm serial attached small computer system interface [SAS][1] drives) mean faster read/write speeds. All things being equal, power use is proportional to the cube of the disk spin speed. To reduce storage energy use, storage administrators should look for slower drives (e.g., 7.5 K rpm serial advanced technology attachment [SATA][2] drives) that are available to accommodate specific tasks at hand.
- **Massive array of idle disks (MAID).** MAID operates more energy efficiently than older systems and often offers an effective solution for Tier 3 storage (data accessed infrequently). MAID saves power by shutting down idle disks, then powering the disks back up only when an application must access the data.
- **Solid-state drives (SSDs).** Energy-saving, solid-state storage increasingly offers an energy-efficient option. Without powering spinning disks, SSDs provide "read" speeds 10 times faster than hard disks. For example, compared to a 7.2 K rpm SATA disk, an SSD consumes one ninth the power per byte stored (Pflueger 2010). SSDs are, however, more expensive than conventional hard disk options.
- **ENERGY STAR-certified data storage (EPA undated b).** EPA's ENERGY STAR program certifies energy-efficient online data storage that meets the following criteria:
    - Employs efficient power supplies that limit power conversion losses.
    - Relies on internal variable-speed fans for cooling.
    - Provides features to help better manage data, leading to reduced storage and energy consumption.

[1] SAS is a faster and historically more expensive interface that moves data to and from storage devices.
[2] SATA is the next-generation computer bus interface that moves data to and from storage devices.

2 Application Conditions of the Protocol

Unlike other efficiency measures in the Uniform Methods Project, data center IT measures present a new target for utility programs.[3] As shown in Table 1, most utilities offer custom incentives for data center IT measures, where applicants must calculate and demonstrate savings from data center IT equipment. Utilities pay incentives based on actual verified savings. Table 1 shows a range of $0.06 to $0.16/kWh saved. In general, standard custom programs work in the following manner:

- A customer submits a project application that includes energy use of existing equipment, equipment required by code or standard, and the efficiency measure (PG&E 2013). In addition, customers must specify whether they install the efficiency measure as an early replacement (where an existing unit has remaining useful life) or at burnout (where the existing unit no longer operates).
- The utility inspects and approves the project before removing the existing equipment/systems and installing the new equipment/systems.
- Upon completion of the project, the utility inspects and approves installation of the measures and finalizes the incentive amounts.

Sometimes utilities offer prescriptive incentives for server virtualization. For example, Seattle City Lights and the Energy Trust of Oregon offer prescriptive incentives based on the number of servers retired. A company in the Seattle City Light territory could receive $900 for retiring six servers through a virtualization effort. In developing the prescriptive incentive, utilities calculated predefined fixed average energy savings, or deemed values, for existing and efficient IT equipment.

Server virtualization also improves scalability, reduces downtimes, enables faster deployments, reduces IT footprints, and has become commonplace, especially in large data centers A 2011 survey of more than 500 large enterprise data centers found that 92% use virtualization to some degree (Veeam 2011). Free-ridership concerns have caused some utilities to remove server virtualization from their data center efficiency programs. Silicon Valley Power's Data Center Program (limited to larger data centers) does not provide incentives for server virtualization. (The program also does not allow IT equipment incentives, unless specifically approved.) PG&E and BC Hydro also stopped offering server virtualization incentives. This trend may continue as organizations redesign data center programs to adjust to market conditions.

[3] As discussed in Considering Resource Constraints in the UMP *Chapter 1: Introduction*, small utilities (as defined under the Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

Table 1. Examples of Data Center IT Incentives Across the Country as of October 2013

Utility	Measure	Incentive Amount	Notes
Seattle City Light (2013)	Custom IT Equipment—Plug Loads	$0.06/kWh saved	Energy savings from custom projects where software or hardware deployments save energy in IT equipment.
	Server Virtualization	$150/server removed	Maximum of 200 servers removed.
NYSERDA (2014)	Examples listed: • Energy-efficient servers, storage, and switches • Server virtualization • Server refresh • Storage consolidation and optimization • High-performance computing systems	$0.12/kWh saved upstate $0.16/kWh downstate	Capped at $5 million per facility.
ComEd (2014)	Examples listed: • Virtualization • Consolidation • Thin-provisioning • Solid state storage	$0.07/kWh saved	Up to 100% of the incremental cost and 50% of the total cost of the project.
Energy Trust of Oregon (2014a, 2014b)	Virtualization	$350 per server decommissioned	10 server minimum
Arizona Public Service (2014)	Example listed: server virtualization	$0.09/kWh	Virtualization listed as "typical custom project," up to 75% of incremental costs.
Southern California Edison (2012)	Reduced process load	$0.08/kWh	Also $100/kW.
Silicon Valley Power (2014a, 2014b)	Virtualization and consolidation of servers, IT equipment	Not Allowed	Large data centers (greater than 350 kW IT load or greater than 100 tons cooling) denied server virtualization/consolidation incentives. General, IT measure savings are not allowed unless specifically approved by SVP.

3 Savings Calculations

3.1 The Simple Algorithm

Unique challenges arise in calculating savings for data center IT measures. On one hand, savings estimates can appear straightforward. For custom incentives, calculations can use data center IT equipment power and energy readings taken from UPS, PDUs, or rack power strips. Estimated energy savings can use power draw readings (in kW) taken before and after measure implementation. Annual savings can be estimated using Equation 1 below:

$$Annual\ Energy\ Savings =$$

$$8760 * (Power\ Draw_{Pre-Existing\ Measure} - Power\ Draw_{Efficient\ Measure}) \tag{1}$$

3.2 Complicating Issues With the Simple Algorithm

A number of challenges can, however, arise when calculating typical energy savings for a data center IT efficiency measure using Equation 1. Figure 1 shows the typical factors involved in calculating early replacement and burnout energy savings for efficiency measures, including power draws (of efficient, standard/code, and preexisting measures) and the useful life (of existing measures and efficiency measures). The challenges include:

- The first challenge (represented by the red circles) arises from the difficulty in determining useful life. IT equipment generally does not stop working: rather, customers replace it for a variety of other reasons. For example, organizations often purchase new servers at the end of the old servers' service agreement or if new server features and capabilities require upgrades. Various International Data Corporation studies indicate organizations replace their servers once every 3 to 5 years (IDC 2010, 2012a, 2012b).
- The second challenge (represented by the blue circles) arises from the varying power draws of IT equipment over time and per business demands, due to changes in the useful work output required of a device (e.g., email server workloads after large-scale layoffs). One would thus ideally normalize energy use for the data center workload to ensure accurate savings estimates. For example, if the data center workload increases just before ENERGY STAR servers are installed, the resulting power draw of the ENERGY STAR servers will be higher, producing underestimated savings. Conversely, if the data center workload decreases before new servers are installed, savings will be overestimated. Many ways to define workload-per-Watt have been proposed and used for data centers (e.g., CPU utilization/Watt, kB transmitted/Watt, GB storage/Watt, various benchmark workloads) (The Green Grid 2009; Pflueger 2010). There is, however, no single metric or industry standard for consistent measurement.

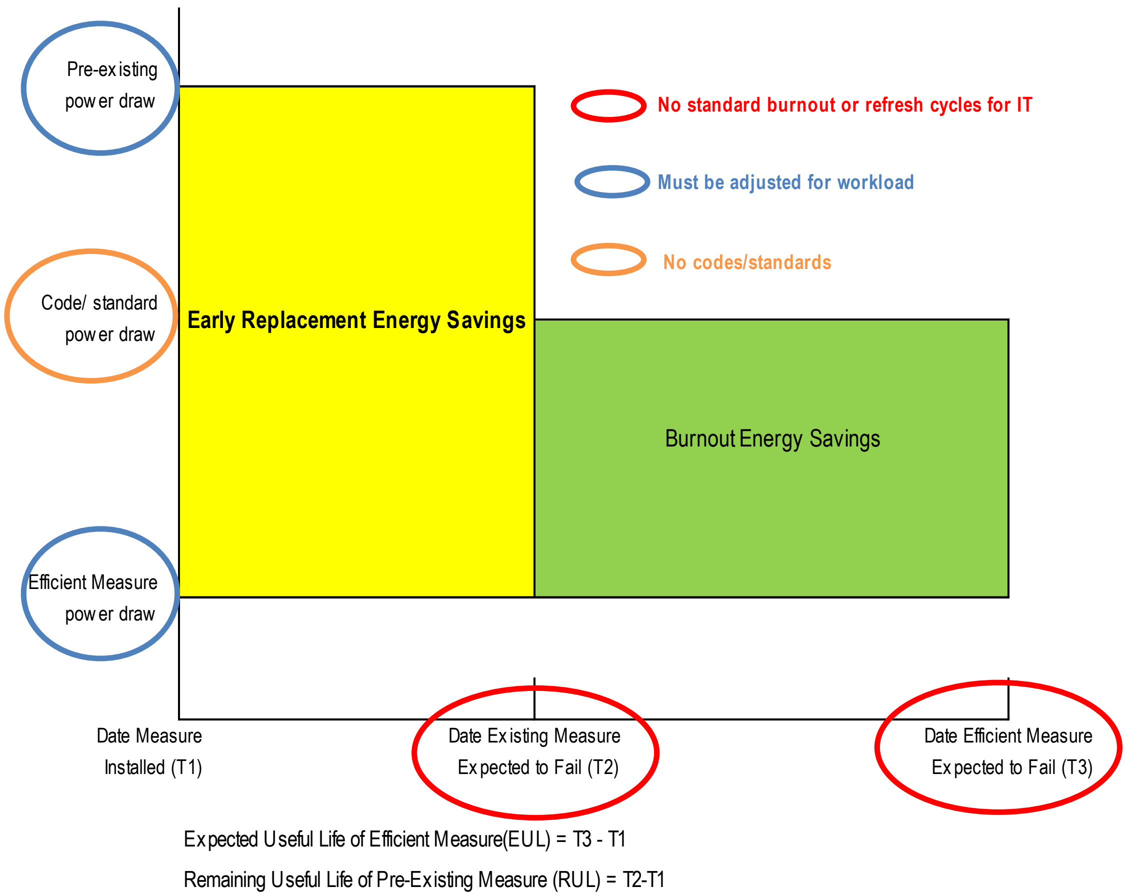


Figure 1. Challenges with determining gross savings of data center IT measures

- The third challenge (represented by the orange circles) arises because—unlike many other efficiency measures in other sectors—energy codes or U.S. Department of Energy standards do not define "typical" or "standard" efficiencies for IT equipment. For such savings estimates, data center operators typically have information about the efficiency measure and preexisting measure, but rarely have information about the "standard" unit, making calculation of burnout savings difficult.

3.3 Calculating Data Center IT Savings

As stated earlier, although in perfect working condition, data center IT equipment often undergoes upgrades when no longer useful (remaining useful life = 0) for reasons other than breaking down (e.g., expired service level agreements, antiquated feature sets, unsatisfactory workload performance issues, incompatibility with hardware-based management systems) (Search Data Center 2012). In other words, "early replacement" savings do not typically apply to data center IT equipment.

Therefore, the following sections present only savings calculations that focus on estimating burnout savings: the energy use difference between the hypothetical "standard" or "typical" equipment available on the market (not the existing equipment) and the efficient equipment to be installed. Figure 2 shows the challenges that remain for calculating the burnout savings of IT equipment.

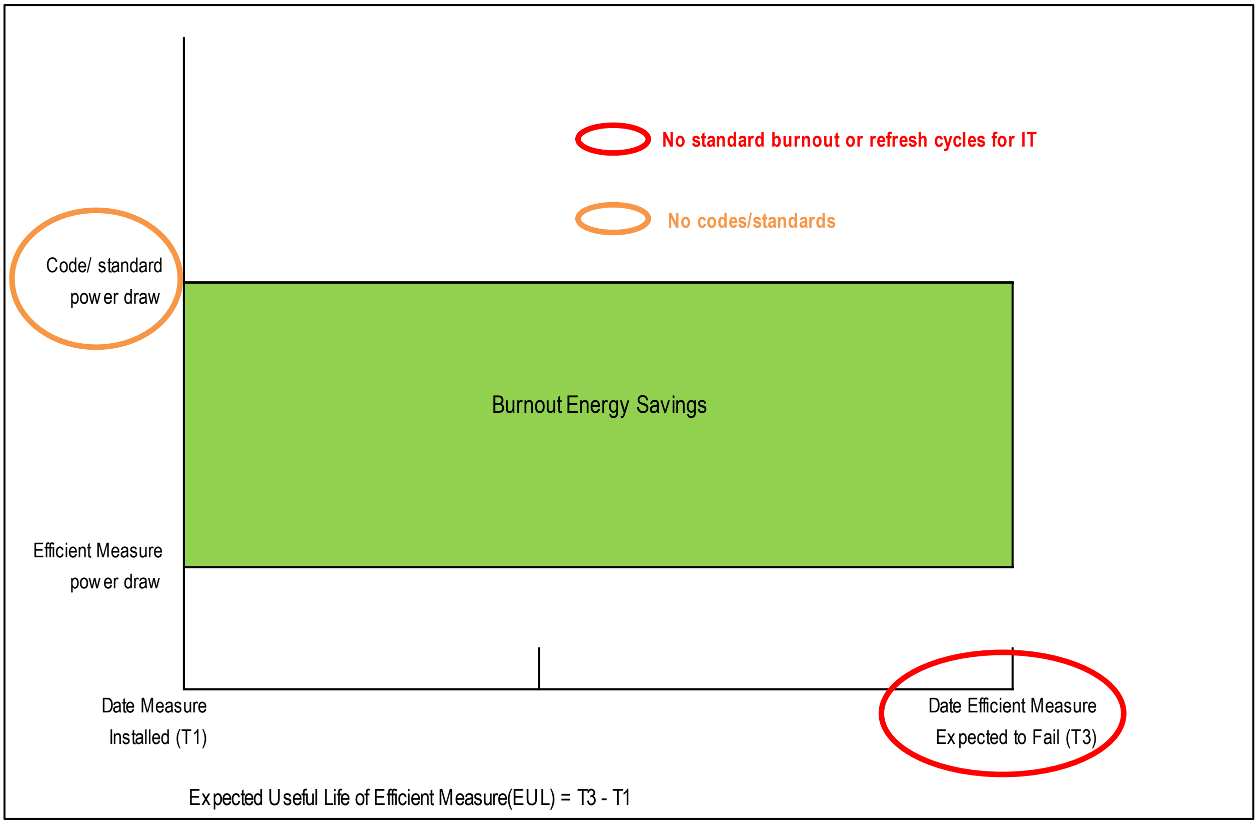


Figure 2. Challenges with determining "burnout only" gross savings of data center IT measures

3.3.1 Calculating Savings When Upgrading to More Efficient Servers

As stated, manufacturers have just started offering server efficiency metrics (EMs) that allow comparisons of server efficiencies.[4] Server EMs soon will allow for simple comparisons between an efficient server and a "baseline" server, which will be established by examining the EMs of servers with similar configurations (e.g., chip sets, memory, and hard drives), computational outputs, and manufacturer years. Equation 2 shows the savings equation when server EMs increase when units becomes more efficient (e.g., operations/Watt), as with the new "efficiency score" generated by SERT. See the Appendix for an example of how the new SERT "efficiency scores" could be used,[5] with Equation 2, to determine the savings from purchasing an energy-efficient server.

$$Annual\ Energy\ Savings_{Efficient\ Servers} = kW_{EE} * (EM_{EE}/EM_{baseline} - 1) * 8760 \qquad (2)$$

[4] EPA requires reporting of the results of SERT, developed by the Standard Performance Evaluation Corporation,
[5] As of October 2014, EPA is just beginning to collect SERT data on servers and has not determined a specific methodology for comparing SERT data at this time.

Where,

kW_{EE}	= power draw in kilowatts of new efficient server equipment
EM_{EE}	= efficiency metric for efficient server
$EM_{baseline}$	= efficiency metric for baseline server
8760	= number of hours in a year as servers run 24/7 in a data center

Another way to calculate savings for servers is to consider ENERGY STAR-certified servers as "efficient servers." Using EPA estimates of percentage savings compared to standard or typical servers, savings can be calculated as shown in Equation 3.

$$Annual\ Energy\ Savings_{ES\ Servers} = (kW_{baseline} - kW_{ENERGY\ STAR}) * 8760$$

$$kW_{ENERGY\ STAR} = \sum_{ES=1}^{n} (kW_{ES,idle} + U_{ES} * (kW_{ES,full\ load} - kW_{ES,idle}) \qquad (3)$$

$$kW_{baseline} = kW_{ENERGY\ STAR}/(1-a)$$

This approach leads to the following simplified expression shown in Equation 4.

$$Annual\ Energy\ Savings_{ES\ Servers} = (\frac{1}{(1-a)} - 1)\ kW_{ENERGY\ STAR} * 8760 \qquad (4)$$

Where,

$kW_{ENERGY\ STAR}$	= power draw in kilowatts of ENERGY STAR server
ES	= ENERGY STAR servers, numbered 1 to n
$kW_{ES,\ idle}$	= power draw in kilowatts of ENERGY STAR server at idle
$kW_{ES,\ full\ load}$	= power draw in kilowatts of ENERGY STAR server at full load
U_{ES}	= utilization of ENERGY STAR server
$kW_{baseline}$	= power draw of baseline servers
a	= percentage ENERGY STAR server is more efficient than baseline "standard" or "typical" unit
8760	= number of hours in a year (servers run 24/7 in a data center)

3.3.2 Calculating Savings for Server Virtualization

Server virtualization savings compare baseline energy use of a large set of single application servers that would have been purchased normally during a server upgrade, without virtualization to a smaller set of virtual host servers, as shown in Equation 5. See the Appendix for an example of how to use SERT data to determine savings from server virtualization.

$$kW_{baseline} = \sum_{1}^{n} (kW_{sa,idle} + U_{sa} * (kW_{sa,full\ load} - kW_{sa,idle}))$$

$$kW_{w\ Virt} = \sum_{1}^{m} (kW_{vh,idle} + U_{vh} * (kW_{vh,full\ load} - kW_{vh,idle})) \qquad (5)$$

$$Annual\ Energy\ Savings_{Virt} = (kW_{baseline} - kW_{w\ Virt}) * 8760$$

Where,

$kW_{baseline}$	= total power draw in kilowatts of all single-application servers without virtualization during server refresh
sa	= single application servers, numbered 1 to n
$kW_{sa,idle}$	= power draw in kilowatts of a single-application server at idle
$kW_{sa,\ full\ load}$	= power draw in kilowatts of a single-application server at full load
U_{sa}	= average utilization of a single-application server over the year
$kW_{w\ virt}$	= total power draw in kilowatts of all virtual hosts
vh	= virtual host servers, numbered 1 to m
$kW_{vh,\ idle}$	= power draw in kilowatts of a virtual host server at idle
$kW_{vh,\ full\ load}$	= power draw in kilowatts of a virtual host server at full load
U_{vh}	= average virtual host server utilization over the year

3.3.3 Calculating Savings for Using More Efficient Storage

Savings from upgrading to more efficient storage equipment (Section 1.4) can be calculated using Equations 6 and 7. Equation 6 uses efficiency metrics of the efficient and baseline unit to estimate savings. Equation 7, similar to Equation 4 (in Section 3.3.1), uses the percentage savings for an ENERGY STAR-certified data storage to estimate savings. To calculate savings from software management tools (Section 1.3), Equation 8 relies on measuring power draws before and after storage management tools are implemented. These power measurements pre- and post-storage management tool should be taken AFTER the efficient storage equipment is installed (if that was also part of the measure) to avoid double counting with savings estimated in Equations 6 and 7.

$$EM_{baseSE} = (\sum_{j=1}^{m} f_{baseSE(j)}\, EM_{baseSE(j)}$$

$$EM_{EESE} = (\sum_{i=1}^{n} f_{EESE(j)}\, EM_{EESE(j)} \tag{6}$$

$$Annual\ Energy\ Savings_{Efficient\ Storage} = kW_{EESE} * (EM_{baseSE}/EM_{EESE} - 1) * 8760$$

Where,

kW_{EESE}	= power draw of new energy-efficient storage equipment
EM_{EESE}	= efficiency metric for energy-efficient storage equipment
EM_{baseSE}	= efficiency metric for baseline storage equipment
$EM_{EESB(j)}$	= Watts per terabyte (TB) of energy-efficient storage device/array j (this value can come from product specifications for devices and/or arrays)

$EM_{baseSB(j)}$	= Watts per TB of baseline device/array j (this value can come from product specifications for devices and/or arrays)
$f_{EESB(i)}$	= fraction of total TB stored on energy-efficient device/array i
j	= baseline devices/arrays, numbered 1 to m
$f_{baseSB(j)}$	= fraction of total TB stored on a baseline device/array j
i	= energy-efficient devices/arrays, numbered 1 to n
8760	= number of hours in a year as servers run 24/7 in a data center

$$Annual\ Energy\ Savings_{ES\ Storage} = \left(\frac{1}{(1-b)} - 1\right) kW_{ES\ STOR} * 8760 \quad (7)$$

Where,

$kW_{ES\ STOR}$	= power draw in kilowatts of ENERGY STAR storage
b	= percentage of ENERGY STAR storage more efficient than typical or standard storage
8760	= number of hours in a year (servers run 24/7 in a data center)

$$Annual\ Energy\ Savings_{\ DS\ Man} = 8760\ * (kW_{Pre\ DS\ Man} - kW_{Post\ DS\ Man}) \quad (8)$$

Where,

$kW_{Pre\ DS\ Man}$	= total power draw in kW of data storage *before* data storage management tool measures implemented (or with tool turned off) and after efficient data storage equipment is installed, if that was part of the measure (the savings from the efficient storage equipment can be calculated using either Equation 7 or 8)
$kW_{Post\ DS\ Man}$	= total power draw in kW of data storage *after* data storage management tools are implemented and after efficient data storage equipment is installed, if that was part of the measure (the savings from the efficient storage equipment can be calculated using either Equation 7 or 8)
8760	= number of hours in a year (servers run 24/7 in a data center)

3.4 Calculating Total Energy, Lifetime, and Peak Demand Savings

Total energy savings, which include additional cooling and power infrastructure savings, can be calculated by multiplying energy savings from an IT upgrade by the data center's power usage effectiveness (PUE):[6] the total data center energy use (e.g., lights; heating, ventilation, and air

[6] This savings calculation assumes the data center's cooling system will be controlled to maintain a given interior temperature set point. When reducing IT power use, less heat must be rejected from the data center. Thus, to maintain a constant temperature set point, cooling system power consumption will be reduced proportional to IT power use reductions. Energy losses at the UPS and transformers also will be reduced proportionally to IT energy-use reductions. Lighting loads may remain constant, but represent only a small fraction of a data center's non-IT energy use. Therefore, PUE remains nearly constant with reduced IT power use. Consequently, total annual energy savings (IT equipment savings plus energy savings in cooling, UPS, and transformer systems) can be reasonably estimated by multiplying PUE by annual IT energy savings.

conditioning [HVAC]; UPS losses; IT) divided by the IT energy use. As a data center becomes more efficient, PUE moves toward 1.

Equation 9 calculates total energy and demand savings.

$$Annual\ Energy\ Savings_{Total} = PUE * Annual\ Energy\ Savings_{IT} \tag{9}$$

Where,

PUE = average PUE determined over the entire year

Equation 10 calculates IT lifetime savings for server virtualization, efficient server upgrades, or efficient storage.

$$Lifetime\ Energy\ Savings_{IT} = Annual\ Energy\ Savings_{Total} * EUL \tag{10}$$

Where,

EUL = expected useful life based on IT upgrade cycle of data center

Equation 11 calculates seasonal peak demand savings, based on server and storage 24/7 operations.

$$Peak\ Demand\ Savings_{Winter} = PUE_{Winter} * Annual\ Energy\ Savings_{IT}/8760$$

$$Peak\ Demand\ Savings_{Summer} = PUE_{Summer} * Annual\ Energy\ Savings_{IT}/8760 \tag{11}$$

Where,

PUE_{Winter} = average PUE over the winter peak demand period,[7] which can be tracked over an entire year. PUE_{Winter} may be smaller in winter due to free cooling)[8]

PUE_{Summer} = average PUE over the summer peak demand period. PUE_{Summer} may be much higher during the summer as free cooling options may not be available as often.

[7] Summer and winter peak demand periods usually vary by state. In Massachusetts, for example, the summer on-peak period is 1:00 pm–5:00 pm on non-holiday weekdays in June, July, and August; the winter on-peak period is 5:00 pm–7:00 pm on non-holiday weekdays in December and January.

[8] Free cooling can include water-side and air-side economization, drastically reducing or eliminating the need for mechanical cooling loads. This is used more often in winter.

4 Measurement and Verification Plan

The following two major savings components must be examined for measures in a data center:

- The power draw of the efficient data center IT equipment.
- The efficiency standards for the measure and for the available IT equipment. (This information allows for development of savings estimates.)

On the surface, the requirements of a typical measurement and verification (M&V) plan for data center IT appear very similar to other energy efficiency measures (e.g., HVAC, lighting). However, given the limited data for EMs in IT spaces and the varied access to data center power draw data, an M&V plan must be flexible and accommodate a wide range of available data.

4.1 International Performance Measurement and Verification Protocol Option

International Performance Measurement and Verification Protocol Option A (Partially Measured Retrofit Isolation) offers the best and only approach for measuring data center IT measures, given its flexibility. Option A relies on field measurements of key performance parameters and estimates of key parameters not selected for field measurements. Data center IT measure energy-use estimates rely on estimates drawn from historical data, manufacturers' specifications, or engineering judgment. Other International Performance Measurement and Verification Protocol options do not provide this flexibility:

- Option B (Retrofit Isolation/Metered Equipment) requires measurement of all energy quantities to compute savings. It does not offer a viable approach because:
    - Data center IT equipment "burnout" savings calculations require using current codes or standards as baseline equipment. As this baseline equipment is not installed, it cannot be metered, and hence cannot fit into an Option B methodology (which requires metering).
    - Generally, a risk-averse manager will not allow metering of IT equipment in a data center. The manager may, however, be able to share data gathered from metering equipment installed at the UPS, PDUs, or in-rack smart power strips.
- Option C uses pre- and post-billing analysis. It also does not present a viable approach. As with Option B, the baseline used in the "burnout" savings calculation draws on current codes or standards, which are not represented in preimplementation electricity bills.

4.2 Verification Process

The verification process involves examining the core assumptions used in developing the savings estimate; this should include the following steps:

- Desk reviews of information pertaining to:
    - Energy-efficient IT equipment
    - Baseline standard or typical IT equipment

    - EMs
    - Efficiency of ENERGY STAR server and storage
    - Power draws
    - EUL
    - PUE
- On-site audits to confirm:
    - Installation of efficient IT equipment
    - Power draws of efficient IT equipment, based on spot readings of UPS, PDU, power strips, and server power
    - Utilization of servers
    - PUE

4.3 Data Requirements/Collection Methods

Table 2 provides details on the types of data needed to verify key inputs for an energy-saving calculation of data center IT equipment, along with methods used for collecting the data:

Table 2. Verification of Key Inputs Into Equations

Key Inputs Into Equations	Verification of Data
Number of Energy-Efficient IT Equipment Units Installed	Reviewers should examine work orders and invoices, and conduct site visits to confirm purchases of efficient units and their installation.
"Baseline" unit	As savings estimates are limited to burnout savings estimates, reviewers should carefully examine how applicants determined baseline standards or typical IT equipment. Baseline IT equipment should: (1) provide the same performance as the energy-efficient it unit (i.e., the same storage capacity in data storage units, same chip set, memory, storage in servers, same computational capacity); and (2) be manufactured in the same year as the energy-efficient IT unit.
Efficiency Metrics for Servers EM_{EE} = efficiency metric for efficient server $EM_{baseline}$ = efficiency metric for baseline server	Reviewers of these metrics should examine SERT. Manufacturers of ENERGY STAR-certified servers must include SERT. Please see the Appendix for an example of how one could interpret and use SERT "efficiency score" data to calculate savings for efficient servers and server virtualization.

Key Inputs Into Equations	Verification of Data
Efficiency Metrics for Storage $EM_{EESB(j)}$ = Watts/TB of energy-efficient storage device/array j (this value can come from product specifications for devices and/or arrays) $EM_{baseSB(j)}$ = Watts/TB of baseline device/array j (this value can come from product specifications for devices and/or arrays)	As shown in Figure 3, the energy use required for data storage varies by technology and disk speed. Energy use can decrease by an order of magnitude with equipment upgrades if an organization replaces faster spinning (15 K rpm) fiber channel hard disc drives (HDDs) with energy-efficient, yet very costly, solid state drives (SSDs). The Storage Networking Industry Association Emerald Power Efficiency effort (http://snia.org/emerald/view) is gathering data on storage device efficiency. In addition to the SNIA, the ENERGY STAR program's new data storage specification, effective on December 2013, has asked data storage makers to provide the following types of performance data for online systems (those with <80ms response time): • Transaction workload (input/output per second {IOPS} per watt through the "Hot Band" and "Random Read/ Write" tests) that mimics a scenario where a large number of random I/O operations are requested with low seek times (e.g., banking); • Streaming workload (MiB per second per watt through the "Random Sequential Read/Write" test) that mimics accessing large continuous chunks of data (e.g., Netflix) and ; • Capacity workload (GB raw capacity per watt through the "Ready Idle" test) that mimics a situation where data is not accessed frequently but must be "ready" (e.g., hospital records). As these data become more readily available for different data storage systems, comparisons of energy efficiency will be possible. For example, as shown in Table 3 below, data from the ENERGY STAR certified data center storage device list shows SSDs to be an order of magnitude more efficient than HDDs for most of the workloads. Note that pure Network Attached Storage (NAS) and tape solutions are not currently covered by this program.
Percent savings for ENERGY STAR IT Equipment *a* = percentage ENERGY STAR server is more efficient than baseline "standard" or "typical" unit *b* = percentage ENERGY STAR storage is more efficient than baseline "standard" or "typical" unit	Reviewers should confirm the estimates for servers and data storage, as provided at the ENERGY STAR website www.energystar.gov/products.

Key Inputs Into Equations	Verification of Data
Power Draws of Servers and Data Storage Based Off Measurements kW_{EE} = power draw of new efficient server equipment $kW_{ENERGY\ STAR}$ = power draw of ENERGY STAR server $kW_{w\ virt}$ = total power draw in kilowatts of all virtual hosts kW_{EESE} = power draw of new energy efficient storage equipment $kW_{ES\ STOR}$ = power draw in kilowatts of ENERGY STAR storage $kW_{Pre\ DS\ Man}$ = total power draw in kW of data storage before data storage management measures implemented $kW_{Post\ DS\ Man}$ = total power draw in kW of data storage after data storage management tools are implemented	Power draw measurements can be taken from: data center energy management systems, storage management tools, UPS, PDUs, power strip with metering capability, or even the actual server or data storage units directly. For example, ENERGY STAR-certified servers must "provide data on input power consumption (W), inlet air temperature (°C), and average utilization of all logical CPUs." (EPA 2013). When examining measured power draw data, reviewers should look to: (1) review data averaged over a month to account for differences in server loads on weekends and nights or in differing storage levels used due to data storage resource management tools; and (2) account for PDU or UPS power losses when measuring IT equipment at the PDU or UPS. Although the data center manager probably will not allow confirmatory metering of power draws of IT equipment, options may be available to meter at electrical panels feeding specific data center loads.
Full Load and Idle Load Power Draws of Servers Based Off Manufacturer's Data $kW_{sa,idle}$ = power draw in kilowatts of a single-application server at idle $kW_{sa,\ full\ load}$ = power draw in kilowatts of a single-application server at full load $kW_{vh,\ idle}$ = power draw in kilowatts of a virtual host server at idle $kW_{vh,\ full\ load}$ = power draw in kilowatts of a virtual host server at full load	Reviewers of these metrics should examine SERT. Manufacturers of ENERGY STAR-certified servers must include SERT data that will include full-load and idle load data. Please see the Appendix for an example of how to interpret and use SERT idle and full-load power draw data to calculate savings for efficient servers and server virtualization.
Utilization of Servers U_{vh} = average virtual host server utilization over the year U_{ES}= utilization of ENERGY STAR server U_{sa} = average utilization of a single-application server over the year	For the installed virtual host server or installed ENERGY STAR server, utilization of servers should be derived from a data center's server performance software. Utilization of a baseline single application server may be estimated based on past implementations before server virtualization was implemented.
EUL	Reviewers should recognize that IT upgrades generally occur every 3 to 5 years, but can vary by organization. IT managers should base EULs on historical data from past hardware purchases and refresh cycles, as those EULs will be much more accurate for a given organization. When such information is not available, an IT manager might use 5 years for smaller data centers and 3 years for larger data centers, based on national average refresh cycles. The reviewer should also ask to compare a recommended EUL to: • Length of data center service-level agreements. • Time period since last IT upgrade.

Key Inputs Into Equations	Verification of Data
Power Usage Effectiveness (PUE) PUE is the total data center energy use (e.g., lights, HVAC, UPS losses, IT) divided by the IT energy use.	Reviewers must recognize that the quality of the PUE estimate varies a great deal across data centers. Larger data centers have an in-house PUE estimates, tracked over time. According to a 2013 recent Uptime Institute industry survey of large data centers, PUE averages roughly 1.65 and 66% of large data centers measure PUE. Google's large chiller-less data centers have achieved a PUE of 1.1 (Miller 2011). Many small data center spaces (e.g., server closets and rooms, localized data centers smaller than 1,000 ft^2) may have never measured PUE. Evidence suggests that in some cases, poorly managed data center cooling, lack of variable-speed fans, load reduction that leads to reduced UPS efficiency, and other issues cause PUE to worsen (rise) after reducing IT load in data centers and not stay constant. Therefore, reviewers are encouraged to use PUE estimates after IT load is reduced. Numerous online models (developed by the Green Grid, APC, and others) exist, some are simple and some relatively complex, to estimate PUE (Karthi 2008). We recommend two guides for measuring PUE: • A multiparty task force (composed of 7x24 Exchange, ASHRAE, The Green Grid, Silicon Valley Leadership Group, U.S. Department of Energy Save Energy Now Program, U.S. Environmental Protection Agency's ENERGY STAR Program, United States Green Building Council, and Uptime Institute) developed a 12-page guidance for measuring and reporting PUE (EPA 2011). • The Green Grid developed an 80-page document titled "PUE: A Comprehensive Examination of the Metric" in 2012. This document supersedes previous white papers and consolidates all information that The Green Grid has developed and published relating to PUE (The Green Grid 2012).

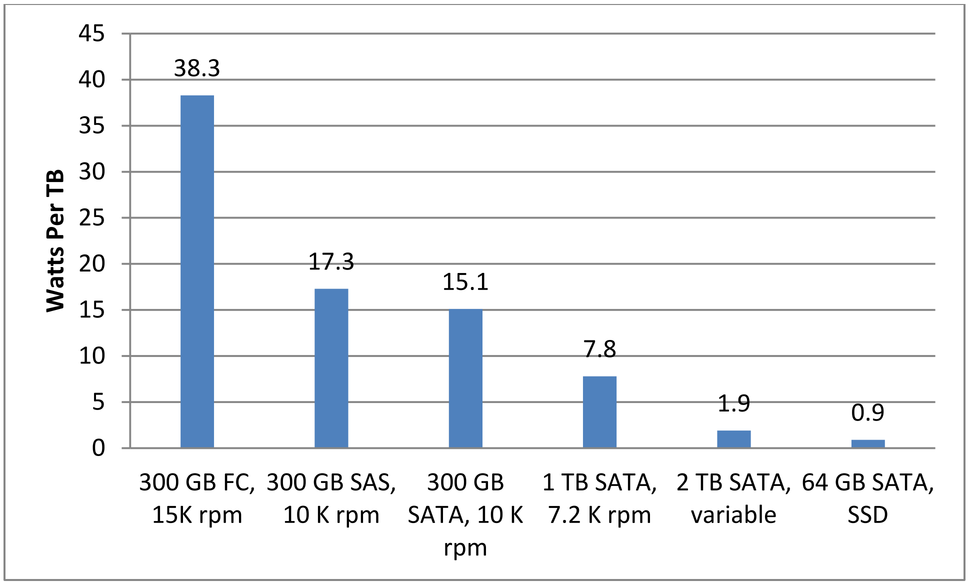


Figure 3. Watts per terabyte for various data storage types (Pflueger 2010)

Table 3. ENERGY STAR-Certified Storage Workload Test Results

Type of Storage	Hot Band Workload Test (IOPS/W)	Random Read Workload Test (IOPS/W)	Random Write Workload Test (IOPS/W)	Ready Idle Workload Test (GB/W)
SSD	138	1069	254	37
HDD	30	18	23	13

5 Other Evaluation Issues

Two issues can complicate evaluation of data center IT equipment savings (EPA 2012):

- **Long lead times.** Data center deployments often take longer to complete than other types of energy efficiency engagements. All projects, whether related to IT equipment or to its supporting infrastructure, require careful planning and execution. These long lead times may complicate evaluating savings, as the project simply may not be completed by the time evaluation takes place. Evaluating savings before completion of an IT upgrade may result in significantly smaller savings than originally estimated.
- **Short production cycles.** Servers and many other types of IT equipment have annual production cycles due to frequent technological upgrades. These production cycles differ from product categories such as HVAC equipment, food service equipment, and residential appliances, which generally advance over multiyear timeframes. Technological advances can cause data center equipment to become antiquated with relative frequency. Thus, savings calculations for IT equipment should be based on a "burnout" scenario, comparing the efficiency measure to the baseline standard or typical equipment available at the time of installation. During the evaluation, reviewers must carefully examine the baseline equipment available at the time of the IT efficiency measure's installation. If the baseline equipment does not represent the equipment available at the time of the efficiency measure's installation, savings could be significantly underestimated, given the short production cycles and how quickly IT equipment efficiency increases over time.

6 References

Arizona Public Service (2014). "Retrofit & New Construction Rebates." http://www.aps.com/en/business/savemoney/rebates/Pages/custom-rebates.aspx

Clark, T.; Yoder, A. (2008). "Best Practices for Energy Efficient Storage Operations Version 1.0." Storage Networking Industry Association's Green Storage Initiative, www.snia.org/sites/default/files/GSI_Best_Practices_V1.0_FINAL.pdf.

ComEd (2014). "Stay Cool While Saving Energy in Your Data Center." https://www.comed.com/business-savings/commercial-industrial/Pages/data-centers.aspx.

DOE (2011). "Energy 101: Energy Efficient Data Centers." U.S. Department of Energy www.youtube.com/watch?v=xGSdf2uLtlo.

Energy Trust of Oregon (2014b). "IT Power: Rack Up the Savings." http://energytrust.org/commercial/incentives/equipment-upgrades-remodels/Software/it-power/;

Energy Trust of Oregon (2014a). "Incentives: Data Centers." http://energytrust.org/library/forms/be_pi0195d.pdf

EPA (undated a). ENERGY STAR "Consolidation of Lightly Utilized Servers." U.S. Environmental Protection Agency, www.energystar.gov/index.cfm?c=power_mgt.datacenter_efficiency_consolidation

EPA (undated b). "ENERGY STAR Server Specifications." U.S. Environmental Protection Agency www.energystar.gov/products

EPA (2011). "Recommendations for Measuring and Reporting Overall Data Center Efficiency." U.S. Environmental Protection Agency, www.energystar.gov/ia/partners/prod_development/downloads/Data_Center_Metrics_Task_Force_Recommendations_V2.pdf?5311-22ed.

EPA (2012). "ENERGY STAR Understanding and Designing Energy-Efficiency Programs for Data Centers." U.S. Environmental Protection Agency, www.energystar.gov/ia/products/power_mgt/ES_Data_Center_Utility_Guide.pdf?ff29-42fa.

EPA (2013). "ENERGY STAR Program Requirements for Computer Servers." Section 5.1. U.S. Environmental Protection Agency, www.energystar.gov/products.

IDC (2012b). "The Cost of Retaining Aging IT Infrastructure." Sponsored by HP. http://mjf.ie/wp-content/uploads/white-papers/IDC-White-Paper_the-cost-of-retaining-aging-IT-infrastructure.pdf

IDC (2010). "Strategies for Server Refresh." Sponsored by Dell. http://i.dell.com/sites/content/business/smb/sb360/en/Documents/server-refresh-strategies.pdf

IDC (2012a). "Analyst Connection: Server Refresh Cycles: The Costs of Extending Life Cycles." Sponsored by HP/Intel. Online. http://resources.itworld.com/ccd/assets/31122/detaill

Karthi (2011). Site 24x7. "Five Online Calculators Available to Calculate Energy Efficiency of a Data Center." https://blogs.site24x7.com/2011/08/26/five-online-calculators-available-to-calculate-energy-efficiency-of-a-data-center/

Koomey, J. (2011). “Growth in Data Center Electricity Use in 2005 to 2010.” www.analyticspress.com/datacenters.html.

Miller, R. (2011). “Chillerless Data Center is Google’s Top Performer.” www.datacenterknowledge.com/archives/2011/05/26/chiller-less-data-center-is-googles-top-performer/.

Netapp (2014). “Simple Provisioning of Your Unified Storage Environment.” www.netapp.com/us/technology/unified-storage/simple-provisioning.aspx

NYSERDA (2014). “Data Center Incentives Through the Industrial Process and Efficiency Program.” https://www.nyserda.ny.gov/Energy-Efficiency-and-Renewable-Programs/Commercial-and-Industrial/Sectors/Data-Centers.aspx.

Pflueger, J. (2010). “Understanding Data Center Energy Intensity.” A Dell Technical White Paper.

Veeam (2011). “Veeam Launches V-Index To Measure Virtualization Penetration Rate.” www.veeam.com/news/veeam-launches-v-index-to-measure-virtualization-penetration-rate.html.

Search Data Center (December 7, 2012). “The server lifecycle: Is it time for that aging server to go?” http://searchdatacenter.techtarget.com/tip/The-server-lifecycle-Is-it-time-for-that-aging-server-to-go.

Seattle City Light (2013) “2013 Energy Conservation Incentives: Medium and Large Commercial.” Online. http://www.seattle.gov/light/Conserve/Business/2013Incentives.pdf

Silicon Valley Power (2014b). “2013–2014 Data Center Program Rebate Application.” http://siliconvalleypower.com/Modules/ShowDocument.aspx?documentid=5561.

Silicon Valley Power (2014a). “2013–2014 Customer Directed Rebate Application.” http://siliconvalleypower.com/Modules/ShowDocument.aspx?documentid=5102.

SPEC (2014). “Server Efficiency Rating Tool Design Document 1.0.2.” Standard Performance Evaluation Corporation, www.spec.org/sert/.

Southern California Edison (2012). “Data Center Energy Efficiency Program.” http://www.willdan.com/energy/DCEEP.aspx.

The Green Grid (2009). “Proxy Proposals for Measuring Data Center Efficiency.” Online. www.thegreengrid.org/Global/Content/white-papers/Proxy-Proposals-for-Measuring-Data-Center-EfficiencyYoder, A. (2012) “Technologies for Green Storage.” Storage Networking Industry Association presentation,” http://snia.org/emerald

The Green Grid (2012). “PUE: A Comprehensive Examination of the Metric.” www.thegreengrid.org/en/Global/Content/white-papers/WP49-PUEAComprehensiveExaminationoftheMetric

7 Appendix: Hypothetical Calculations of Savings From an Efficient Server or Server Virtualization Using SERT Data

SERT was created by Standard Performance Evaluation Corporation to measure server energy efficiency by using a set of synthetic[9] workloads or worklets as they are called, to test discrete system components such as processors, memory and storage, providing detailed power consumption data at different load levels. SERT, required by the ENERGY STAR program, allows a data center manager to compare server "efficiency score" and power draw across various workload types and at various loads between two or more candidate systems. As SERT becomes more widely adopted, it will allow for more standardized energy savings determinations. We provide some background on SERT and the manner in which savings could be determined. The worklets are grouped into four workload types shown in Table 4:

Table 4. SERT Workload Types, Worklet Names, and Characteristics

(SPEC 2014)

Workload Type	Worklet Names	Characteristics
CPU	Compress, CryptoAES, LU, SHA256, SOR, SORT, XML Validate	• The worklet requires consistent processor characteristics per simulated "user" regardless of number of processors, cores, enabled threads, etc. • At the 100% load level, the performance bottleneck is the processor subsystem. • The worklet's performance should increase with more processor resources, including the number of processors, the number of cores, possibly the number of logical processors, increased frequency, larger available cache, lower latency, and faster interconnect between CPU sockets. • Readings at loads of 25%, 50%, 75%, and 100%.
Memory	Flood, Capacity	• The worklet contains consistent memory access characteristics per simulated "user" regardless of size and number of dynamic inline memory modules. • At the 100% load level, the performance bottleneck is the memory subsystem. • The worklet's performance should measure a higher (better) performance score with improved memory characteristics (e.g., higher bandwidth, lower latency, total memory size). • The worklets as a group should reflect a combination of random and sequential reads and writes, and small and large memory accesses. • Readings at loads of 50% and 100% for Flood. • Readings at 4, 8, 16, 128, 256, 512, 1024 GB for Capacity.

[9] Synthetic workloads or worklets are discrete operations of a specific type that are repeated over and over again. They represent theoretical capabilities of the system that are rarely exercised in such a repetitious and discrete manner in the real world. Application benchmarks more represent typical activities but don't allow specific performance capabilities to be isolated.

Workload Type	Worklet Names	Characteristics
Storage I/O	Random, Sequential	• The worklets reflect consistent input/output characteristics per simulated “user” regardless of system size and number of disks or the installed memory. • The worklets consist of a combination of random and sequential accesses, reads and writes, and small and large inputs and outputs. • At the 100% load level, the performance bottleneck is the storage subsystem. • The worklets should score a higher (better) performance result for higher bandwidth and lower latency. • The worklets are limited to testing individual internal storage devices only. RAID arrays and external storage devices are not supported. • Readings at loads of 50% and 100%.
Hybrid	SSJ	• The worklet reflects a combination of a wide variety of processor and memory-intensive tasks. • At the 100% load level, the performance bottleneck is due to multiple subsystems. • The combined worklets should measure a higher (better) performance score for improved processor and memory characteristics. • Readings at loads of 12.5%, 25%, 37.5%, 50%, 62.5%, 75%, 87.5%, 100%.
Idle	Idle	No transactions occur during this measurement although the server is in a state in which it is capable of completing a transaction

Figure 4 and 5 show SERT sample outputs on two hypothetical servers – Server A and Server B. (These data do not represent any particular server models.) The “efficiency score” column is the “normalized performance” divided by “Watts” across different loads. Idle power draws are indicated by the vertical blue line in the “Watts” column. Hash marks along each horizontal line represent data at the load levels specified in Table 4 for each worklet.

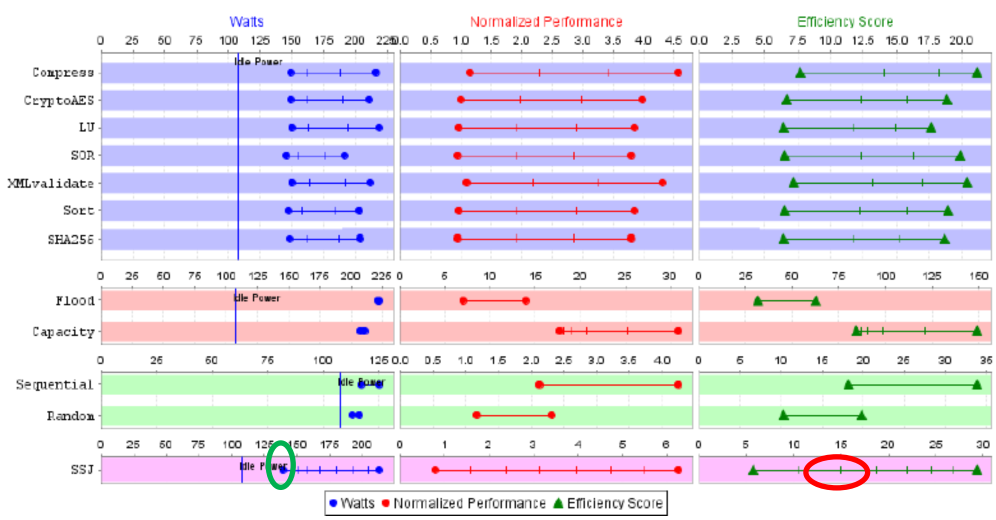


Figure 4. Sample SERT data for Server A

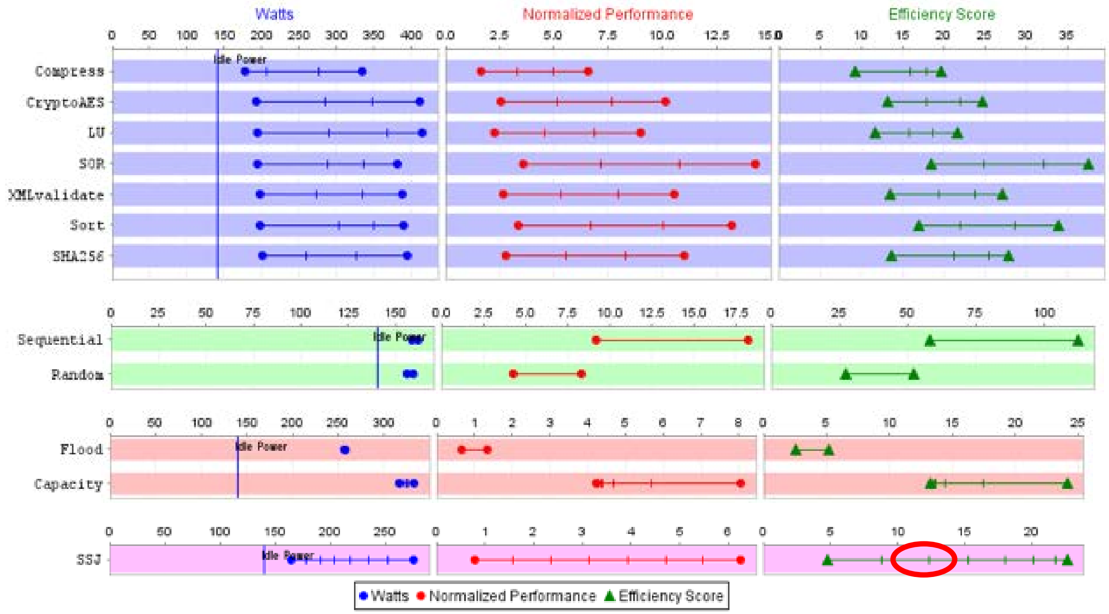


Figure 5. Sample SERT data for Server B

7.1 Savings From an Efficient Server

More information will be available in the near future from EPA and manufacturers in the next year or so about how SERT should be used to compare servers. Until this information is available, to calculate savings from purchasing the Server A instead of Server B, Equation 2 (from Section 3.3.1) could be used to take advantage of the availability SERT efficiency metrics:

$$Annual\ Energy\ Savings_{Efficient\ Servers} = kW_{EE} * (EM_{EE}/EM_{baseline} - 1) * 8760$$

Where,

kW_{EE}	= power draw of new efficient Server A (kilowatts)
EM_{EE}	= efficiency metric (SERT efficiency score) for efficient Server A
$EM_{baseline}$	= efficiency metric (SERT efficiency score) for baseline Server B
8760	= number of hours in a year as servers run 24/7 in a data center

The steps to be taken include:

- **Determine which workload is most appropriate.** SERT has different workload types for CPU, memory, and storage-intensive loads. If you are unclear what type of load is present, you can use as an exemplar, the Hybrid SSJ worklet. This workload reflects a combination of synthetic loads to a wide variety of processor and memory-intensive tasks.
- **Determine the appropriate baseline model.** During the purchase of Server A, other alternative servers, such as Server B, were examined with similar CPU, memory and storage capacity. Server B was selected as the baseline model.

- **Measure the wattage of the efficient Server A.** Using the data center infrastructure management system data, the average power draw of the efficient Server A, kW_{EE}, is **160 Watts**.
- **Estimate the percentage load on the efficient server.** Using the data center's server performance software, utilization is estimated at **25%.**
- **Determine the efficiency scores at the appropriate worklet and load level.** In our example, the SERT efficiency scores for the SSJ worklet at 25% load were determined to be **15** for the efficient Server A (EM_{EE}) and **12.5** for the baseline Server B ($EM_{baseline}$). (See red circles on Figures 4 and 5.)

Using the values determined above, annual energy savings for an efficient server are estimated as:

$$Annual\ Energy\ Savings_{Efficient\ Servers} = 280\ kWh$$
$$= 0.16\ kW * \left(\frac{15}{12.5} - 1\right) * 8760hr$$

7.2 Savings From Server Virtualization

The example below demonstrates using SERT data to estimate savings from server virtualization. In order to calculate savings from server virtualization, Equation 5 (in Section 3.3.2) would normally be used:

$$kW_{baseline} = \sum_1^n (kW_{sa,idle} + U_{sa} * (kW_{sa,full\ load} - kW_{sa,idle}))$$
$$kW_{w\ Virt} = \sum_1^m (kW_{vh,idle} + U_{vh} * (kW_{vh,full\ load} - kW_{vh,idle}))$$
$$Annual\ Energy\ Savings_{Virt} = (kW_{baseline} - kW_{w\ Virt}) * 8760$$

Where,

$kW_{baseline}$	= total power draw in kilowatts of all single-application servers (assumed Server A) without virtualization during server refresh
sa	= single application servers, numbered 1 to n
$kW_{sa,idle}$	= power draw in kilowatts of a single-application server (assumed Server A) at idle
$kW_{sa,\ full\ load}$	= power draw in kilowatts of a single-application server (assumed Server A) at full load
U_{sa}	= average utilization of a single-application server over the year
$kW_{w\ virt}$	= total power draw in kilowatts of all virtual hosts
vh	= virtual host servers, numbered 1 to m
$kW_{vh,\ idle}$	= power draw in kilowatts of a virtual host server at idle
$kW_{vh,\ full\ load}$	= power draw in kilowatts of a virtual host server at full load
U_{vh}	= average virtual host server utilization over the year

Server A, depicted in Figure 4, was assumed to represent the baseline single application server. In addition, it was assumed the baseline scenario is 20 single application server, n = 20, and the virtualization scenario uses two virtual host servers, m = 2. Because of the available metering data and SERT data of power draws at different loads, the equations relying on idle and full-load power draws to estimate savings are not necessary. Instead, the steps to be taken include:

- **Determine which workload is most appropriate.** Since it is unclear what type of load will be present, a Hybrid SSJ worklet, which reflects a combination of loads, is selected.
- **Estimate the wattage of the installed virtual hosts.** Using the DCIM system data, the average power draw of the two virtual host servers, $kW_{w\,virt}$, is **400 Watts**.
- **Estimate the wattage of the single application servers that would have been purchased.** For the baseline estimate, the alternative scenario is a conventional server upgrade where 20 single old application servers were replaced with 20 new single application servers (assumed to be Server A). The average utilization for the single application servers, U_{sa}, was assumed to be **12.5%,** based on IT manager estimates of the load on single application servers run in the past. As shown in the green circle in Figure 4, using the Hybrid SSJ worklet, the wattage at 12.5% load for Server A is **140 Watts**.

Using the values determined above, annual energy savings for a virtualization effort are estimated as:

$$kW_{baseline} = 2.8\ kW = \ 20 * 0.14\ kW$$

$$kW_{w\ Virt} = 0.40\ kW$$

$$Annual\ Energy\ Savings_{Virt} = 21{,}240\ kWh = (2.8\ kW - 0.40\ kW) * 8760$$

NREL
NATIONAL RENEWABLE ENERGY LABORATORY

Chapter 21: Estimating Net Savings – Common Practices

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in September 2014. The content in this version has been updated.

Daniel M. Violette, Ph.D.
Navigant
Boulder, Colorado

Pamela Rathbun
Tetra Tech
Madison, Wisconsin

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68578
October 2017

Contract No. DE-AC36-08GO28308

Chapter 21: Estimating Net Savings – Common Practices

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in September 2014. The content in this version has been updated.

Daniel M. Violette, Ph.D.
Navigant
Boulder, Colorado

Pamela Rathbun
Tetra Tech
Madison, Wisconsin

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68578
October 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter authors wish to thank and acknowledge the Uniform Methods Project Steering Committee and Net-to-Gross Technical Advisory Group members for their contributions to the original Net Savings chapter and/or this updated version of the chapter. The following people offered valuable input to the development of this chapter by providing subject-related materials, in-depth discussion, and careful review of draft versions:

- Michael Li, U.S. Department of Energy
- Chuck Kurnik, National Renewable Energy Laboratory
- Michael Rufo, Itron, Inc.
- Hossein Haeri, M. Sami Khawaja, and Arlis Reynolds, The Cadmus Group, Inc.
- Tom Eckman, Northwest Power Planning Council
- Elizabeth Titus, Northeast Energy Efficiency Partnerships
- Steve Schiller, Schiller Consulting
- Rick Ridge, Ridge & Associates
- Ralph Prahl, Ralph Prahl & Associates
- Jane Peters and Marjorie McRae, Research Into Action, Inc.
- Frank Stern, Ken Seiden, and Jeff Erickson, Navigant
- Lynn Hoefgen, NMR Group, Inc.
- Nick Hall, TecMarket Works
- Miriam Goldberg, DNV GL
- Peter Miller, Natural Resources Defense Council
- Dick Spellman, GDS Associates, Inc.
- Tim Xu, Pacific Gas and Electric Company
- Rafael Friedmann, Pacific Gas and Electric Company
- Ryan Firestone, Consultant to the Regional Technical Forum (RTF), Northwest Power and Conservation Council (NWCouncil)
- Christian Douglass, Consultant to the Regional Technical Forum (RTF), Northwest Power and Conservation Council (NWCouncil)
- Bill Provencher, University of Wisconsin (University of Wisconsin, Agricultural and Applied Economics)
- Teri Lutz, formerly of Tetra Tech, made substantive contributions across the entire chapter.

Suggested Citation

Violette, Daniel M.; Rathbun, Pamela. (2017). *Chapter 21: Estimating Net Savings – Common Practices: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68578. http://www.nrel.gov/docs/fy17osti/68578.pdf

Acronyms

C&I	commercial and industrial
CFL	compact fluorescent lamp
DiD	difference-in-differences
EE	energy efficiency
EM&V	evaluation, measurement, and verification
FR	free-ridership
FOE	Focus on Energy
HER	Home Energy Report
IOU	investor-owned utility
ISP	Industry Standard Practice
kWh	kilowatt-hours
LFER	linear fixed-effects regression
MCM	macroconsumption metric
ME	market effects
NPSO	nonparticipant spillover
NTG	net-to-gross
NW Council	Northwest Power and Conservation Council
PSO	participant spillover
RCT	randomized control trial
RDD	regression discontinuity design
RED	random encouragement design
RTF	Regional Technical Forum
SRA	self-report approach
TRM	technical reference manual
UMP	Uniform Methods Project

Protocol Updates

The original version of this protocol was published in September 2014.

This chapter has been updated to incorporate the following revisions:

- Modified the definitions of net and gross savings
- Reorganized the chapter slightly by:
    - Dividing the section on experimental design into two separate sections—one focusing on approaches that use random assignment (e.g., randomized control trials) and a second addressing quasi-experimental design approaches
    - Adjusting the order in which methods are presented to improve the logical flow of the chapter
- Expanded the discussion of survey methods based on recent developments in the literature
- Updated the Common Practice Baseline section with examples of how they have been set.

Table of Contents

List of Tables

1 Estimating Net Energy Savings

This chapter focuses on the methods used to estimate net energy savings in evaluation, measurement, and verification (EM&V) studies for energy efficiency (EE) programs. The chapter provides a definition of net savings, which remains an unsettled topic both within the EE evaluation community and across the broader public policy evaluation community, particularly in the context of attribution of savings to a program. The chapter differs from the measure-specific Uniform Methods Project (UMP) chapters in both its approach and work product. Unlike other UMP resources that provide recommended protocols for determining gross energy savings, this chapter describes and compares the current industry practices for determining net energy savings but does not prescribe methods.

Readers should treat this chapter as a resource document that provides state-of-the-art information about common practices for determining net energy savings. The selection and description of methods are based on the results of research by EM&V experts. The chapter describes the common methods and the approaches that are receiving attention in the evaluation community and discusses how net savings values are used for reporting and for energy-system planning.

The determination of net savings is an issue in EE programs that are funded publicly or through utility-customer resources. For these programs, the most direct contribution of net savings evaluation studies is to provide decision-makers the information they need to make good EE investments. Program goals, scale, funding sources, and the specific audience for the evaluation effort can influence the methods used, the aspects of the evaluation that are emphasized, the depth of analysis, and the way the results are presented.

Estimating net savings is central to many EE evaluation efforts and is broad in scope. It requires the determination of baselines (i.e., the counterfactual) and savings levels across many types of programs. The intent of this document is to present information on the tradeoffs in the various methods for calculating net savings that will help policy-makers, regulators, and program administrators decide which are best to apply.

The references section at the end of this chapter includes cited articles that address the presented methods in greater depth than the scope of this chapter allows.

2 Universality of the Net Impacts Challenge

Investment decisions result in allocating resources to achieve objectives. Regardless of the type of investment, once made, it is difficult to assess what would have happened absent that decision. This is the essence of evaluation: "What are the impacts of that investment decision?" These are termed *net impacts*, or *attributable impacts*. To address net impacts, a baseline is needed that represents what would have happened in the absence of the investment. This baseline is also called the *counterfactual scenario*.[1]

The broader literature on evaluation reveals a parallel between issues arising from estimating the net impacts of EE investments and estimating the effects of other types of investments made in either the private or the public sector. Examples include:

- Healthcare: What would the health effects have been without an investment in water fluoridation?
- Tax subsidies for economic development: Would the project—or a variant of the project—have proceeded without a subsidy?
- Education subsidies: What would happen if school lunch programs were not subsidized or if low-interest loans for higher education were not offered?
- Military expenditures: What would have happened without an investment in a specific military program or technology?

Across industries and applications, program evaluators grapple with how to appropriately approximate the counterfactual scenario and determine impacts that are attributable to the investment being analyzed (Cook et al. 2010). [2]

[1] As discussed in the "Considering Resource Constraints" section of the UMP *Chapter 1: Introduction*, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.
[2] Some evaluators also view net savings estimation as an assessment of causality. This chapter uses the term *attribution* rather than *causality*, as it is more descriptive of the problem discussed, whereas causality has a wider range of interpretations that extends to metaphysics.

3 Defining Gross and Net Savings for Practical Evaluation

This section defines key terms related to estimating net savings and summarizes various uses of their measurement in the industry. It also describes many issues evaluators face when estimating net savings in the context of developing an appropriate baseline against which program accomplishments are compared to estimate net impacts.

3.1 Definition of Gross and Net Savings

The following definitions of gross and net savings are used in this chapter:

- **Net savings:** The difference in energy consumption *with the program in place* versus what consumption would have been *without the program in place*.
- **Gross savings:** The difference in energy consumption *with the energy-efficiency measures promoted by the program in place* versus what consumption would have been *without those measures in place*.

3.2 Definitions of Factors Used in Net Savings Calculations

The factors most often considered in net savings calculations are free-ridership, spillover (both participant and nonparticipant), and market effects. The definitions of these factors shown in Section 3.2.1 and Section 3.2.2 are consistent with those contained in the Energy Efficiency Program Impact Evaluation Guide (SEE Action 2012b). Not all net estimation methods require the explicit estimation of these factors, but they are useful considerations when assessing how completely different estimation methods address net savings in the context of attribution.

3.2.1 Free-Ridership

Free-ridership is the program savings attributable to free-riders (program participants who would have implemented a program measure or practice in the absence of the program). There are three types of free-riders:

- **Total free-riders:** Participants who would have completely replicated the program measure(s) or practice(s) on their own and at the same time in the absence of the program.
- **Partial free-riders:** Participants who would have partially replicated the program measure(s) or practice(s) by implementing a lesser quantity or lower efficiency level.
- **Deferred free-riders:** Participants who would have completely or partially replicated the program measure(s) or practice(s) at a time after the program timeframe.

3.2.2 Spillover

Spillover refers to additional reductions in energy consumption or demand that are due to program influences beyond those directly associated with program participation. As a result, these savings may not be recorded in the program tracking system and credited to the program. There are generally two types of spillover:

- **Participant spillover:** This represents the additional energy savings that are achieved when a program participant—as a result of the program's influence—installs EE measures or practices *outside* the efficiency program after having participated.

  Evaluators have further defined the broad category of participant spillover into the following subcategories:

  - *Inside spillover:* Occurs when participants take additional program-induced actions at the project site
  - *Outside spillover:* Occurs when program participants initiate actions that reduce energy use at sites that are not participating in the program
  - *Like spillover:* Refers to program-induced actions participants make outside the program that are of the same type as those made through the program (at the project site or other sites)
  - *Unlike spillover:* Refers to EE actions participants make outside the program that are unlike program actions (at the project site or other sites) but that are influenced in some way by the program
- **Nonparticipant spillover:** This represents the additional energy savings that are achieved when a nonparticipant implements EE measures or practices as a result of the program's influence (for example, through exposure to the program) but is not accounted for in program savings.

3.2.3 Market Effects

Market effects refer to "a change in the structure of a market or the behavior of participants in a market that is reflective of an increase in the adoption of energy efficiency products, services, or practices and is causally related to market intervention(s)" (Eto et al. 1996). For example, programs can influence design professionals, vendors, and the market (through product availability, practices, and prices), as well as influence product or practice acceptance and customer expectations. All these influences may induce consumers to adopt EE measures or actions (Sebold et al. 2001).[3]

Some experts suggest that market effects can be "best viewed as spillover savings that reflect significant program-induced changes in the structure or functioning of energy efficiency markets." Prahl et al. (2013) also suggest that market transformation is a subset of market effects

[3] When assessing EE policies in a broad context, it should be acknowledged that some participants identified as free-riders in a current program might not have had the opportunity to adopt the EE measure or service were it not for the effects on the market from previous EE program efforts. These efforts may have contributed to that measure or service being available to customers in the current year. The importance of this issue to evaluation depends on the parameters of the evaluation. Most evaluations focus on set time periods spanning 1–3 years. Factors that are included are based on the incremental actions taken as a result of the EE program year being evaluated and the current state of the EE market. Actions taken that resulted from EE efforts in preceding years represent sunk costs and are not incremental to the current program being evaluated. However, this may be an important consideration in a broader policy assessment examining the overall trend in the adoption of EE measures and services across a longer time period. Market effects of previous years' programs may not have been fully accounted for, and this can be a consideration in the broader policy context. However, for assessing the impacts of a given EE program for a given year, these effects from past programs are not generally considered. This is discussed in more detail in Section 4.3.

(as the substantive and long-lasting effects). This view implies that market effects are a subset of spillover. Although spillover and market effects are related, the methods used to quantify these two factors generally differ. Therefore, this chapter addresses them separately.

3.2.4 Net Savings Equations

Evaluators use different factors to estimate net savings for various programs and jurisdictions depending on how a jurisdiction views equity and responsibility (NMR Group, Inc. and Research Into Action 2010). For example, some jurisdictions include only free-ridership in the calculation of net savings; others include both free-ridership and spillover. Some jurisdictions estimate net savings without measuring free-ridership or spillover (market-level estimates of net savings). Messenger et al. (2010) also discuss differences across jurisdictions in the reporting of gross and net savings.

A practitioner who is trying to develop methods to estimate values for these factors will find the definitions provided in this section useful. However, the evaluator must work with the information available, which starts with the tracking system.[4] Evaluators typically view the data in the tracking system as the initial estimate of gross savings. Because free-ridership, spillover, and market effects are untracked values, evaluators should estimate or account for them outside of the program tracking system.[5] A practical way to understand these values is to consider spillover and market effects as savings that are attributable to the program, but that are not included in the program tracking system. Free-ridership represents savings included in the program tracking system that are not attributable to the program.

To estimate net savings, the evaluator first estimates free-ridership, spillover, and market effects, then makes appropriate adjustments to the values in the tracking database (or validated tracking database) as illustrated in Equation 1.[6]

---

[4] The definitions for *free-ridership*, *spillover*, and *market effects* should be integrated with (1) how the utility tracks actual program participation data; and (2) how the utility records information about expected program impacts in the program tracking system. In general, the initial gross savings estimate (in terms of expected energy savings by participant or measure) comes from the tracking system. These data may include "deemed values" negotiated by the stakeholders. These deemed values may include factors that lower the savings of a measure, based on assessments of current practice, codes and standards, and other factors that may directly or indirectly influence how the estimated gross savings are adjusted to estimate net savings. It is important to understand how the gross savings are estimated by project and by participant. In fact, the first recommendation of NMR Group Inc. and Research Into Action (2010) is that the Northeast Region needs a process leading to the development of a consistent definition of *adjusted gross savings*.

[5] Direct estimation methods are available to address free-ridership, spillover, and market effects without estimating each separately. This chapter addresses randomized control trials, quasi-experimental designs, and common practice baselines, each of which essentially is used to adjust the savings estimates in the program tracking system.

[6] A *validated tracking database* is simply a reviewed program tracking database. A review of the tracking database can determine obvious errors, whether adjustments can make the claimed (*ex ante*) savings entries more accurate, and whether any deemed savings values include adjustments that account for net savings factors (for example, an adjusted baseline that captures market trends). The validated tracking system then contains the most accurate information on claimed savings for each participating site or project. The benefits of improved information in the tracking system are discussed by Violette et al. (1993).

Equation 1. Net Savings Including Free-Ridership, Spillover, and Market Effects

Net Savings = Gross Savings – FR + SO + ME not already captured by SO

Where:

FR = free-ridership savings

SO = spillover savings

ME = market effects savings not already captured by SO

In much of the literature, the program evaluation approach involves a net-to-gross (NTG) ratio for which free-ridership, spillover, and market effects are expressed as a ratio to gross savings (Equation 2). These widely used ratios work well for some types of evaluation efforts (for example, survey-based estimations). The term is almost synonymous with estimating net savings and is commonly defined as the ratio of NTG savings for the sample. The population gross savings is then multiplied by the NTG ratio to estimate population net savings.

Equation 2. Net-to-Gross Ratio

NTG Ratio = 1 – FR ratio + SO ratio + ME ratio (where the denominator in each ratio is the gross savings)

When using the NTG ratio defined by specific free-ridership, spillover, and market-effect factors (or ratios), evaluators use Equation 3 to calculate net savings:

Equation 3. Net Savings Calculation Using the Net-to-Gross Ratio

Net Savings = NTG Ratio * Gross Savings

These definitions are essentially standard in the evaluation literature;[7] however, a given jurisdiction may decide not to include free-ridership, spillover, or market effects to estimate net savings. For example, evaluators almost always include free-ridership, but, because of policy choices made in a jurisdiction, most do not always fully consider spillover and market effects (see NMR Group, Inc. and Research Into Action 2010; NEEP 2012). Most evaluators agree that spillover and market effects exist and have positive values, but determining the magnitudes of these factors can be difficult. Increasingly, the trend is to include estimates of spillover in net savings evaluations. The inclusion of market effects is also increasing, but to a lesser degree than spillover. Methods are available to address spillover and market effects and, because there is no debate about whether they exist, these factors should be addressed or considered when estimating net savings. For market effects, evaluators must first conduct a basic assessment as to whether market effects are likely, as well as consider the cost and value of a study. It is important to

[7] Other factors (sometimes called *net impact factors*) are generally considered as adjustments to gross impact estimates. These include rebound, snapback, and persistence of savings. Violette (2013) addresses these factors. As with other NTG factors, evaluations do not treat net impact factors consistently in gross impact calculations, and do not consistently adjust program gross impacts to calculate to a final net impacts number.

know the potential sizes of spillover and market effects for a given program or portfolio so appropriate policy decisions can be made about EE investments.

3.3 Uses of Net Savings Estimates in the Energy Efficiency Industry

Many regulatory jurisdictions discuss the appropriate use of net savings estimates. This is due in part to: (1) the cost of the studies to produce these estimates and (2) a perceived lack of confidence in the resulting estimates.[8] However, evaluators and regulators recognize the advantages of consistently measuring net savings over time as a key metric for program performance (Fagan et al. 2009).

Evaluators generally agree that net savings research can be useful for (SEE Action 2012a, 2012b):[9]

- Gaining a better understanding of how the market responds to the program and using that information to modify the program design (including eligibility and target marketing and incentive levels).
- Gleaning insight into market transformation over time by tracking net savings across program years and determining the extent to which free-ridership and spillover rates have changed over time. This insight might be used to define and implement a program exit strategy.
- Informing resource supply and procurement plans, which requires an understanding of the relationship between efficiency levels embedded in base-case load forecasts and the additional net reductions from programs.
- Assessing the degree to which programs effect a reduction in energy use and demand (net savings is one program success measure that should be assessed).

With respect to the last bullet, Schiller (SEE Action 2012b, pp. 2-5) also discusses the importance of consistently measuring savings across evaluation efforts and having consistent evaluation objectives. For example, evaluators in different jurisdictions assess the achievement of goals and targets as measures of overall EE program performance using different measures of savings: gross savings, net savings, or a combination of the two. There are also differences across jurisdictions where the measure of EE program success is used for calculating financial incentives. There are arguments for basing financial incentives on net savings, as well as arguments for basing incentives on gross savings or a combination of the two.[10]

---

[8] Several experienced evaluators indicated in comments on earlier drafts of this chapter that in their experience, the required level of confidence and precision for estimates of net impacts within the EE field is generally greater than that used in other fields faced with similar types of questions and tradeoffs. The authors generally agree with this observation, but no meta-study comparing target levels of confidence and precision for EE program evaluation with similar evaluations in other fields has been conducted.

[9] Other methods that can and should be used to inform program design and understand market response include process evaluations and market assessments.

[10] As more jurisdictions begin to consider the delivery of EE programs as a business process that requires an investment of resources, they are considering the return on investment (more commonly termed *incentives*), which is typically coupled with performance targets. Jurisdictions can base targets on reaching a certain level of gross savings

A recent NEEP document (NEEP 2016) provides six core principles to inform states' decision regarding their applications of gross and net savings based on policy goals. These six principles are: (1) establish a common understanding, (2) align methods and use with policies, (3) address the value of information from evaluation, (4) apply the concept of symmetry,[11] (5) ensure transparency, and (6) acknowledge multiple views across stakeholders.

3.4 The Net Savings Estimation Challenge—Establishing the Baseline

This chapter discusses estimation methods that rely on the development of a baseline (the assumed counterfactual scenario). This baseline is used to measure the net impacts of a program. If evaluators could identify a "perfect baseline" (i.e., a counterfactual scenario that exactly represents what would have happened if the EE program had not been offered), most of the issues associated with estimating net impacts would not arise.

The evaluator is faced with the challenge of identifying a method that produces a baseline that best represents the counterfactual scenario—in other words, what the participant group (and the market) would have done in the absence of the program.[12] To understand and defend the selection of a method for estimating net savings, the evaluator should consider the implicit and explicit assumptions used for the baseline. For example, when considering the use of nonparticipants as a candidate baseline, the evaluator needs to account for issues that pertain to the similarity, or matching, of the program participants with customers that may comprise the nonparticipant comparison group. The evaluator should also account for any effects the program might have had on the comparison group that may influence the program net savings.

Self-selection can be viewed as a baseline issue that arises when a program is voluntary and participants select themselves into the program, suggesting the potential for systematic differences between program participants and nonparticipants. This issue is not unique to EE evaluations and arises in any policy or program assessment involving self-selection. Specifically, the assumption in this case is that the self-selected participants are those who would have taken more conservation actions than the general nonparticipant comparison group.[13]

---

or on achieving a certain level of net savings—each has pros and cons. A gross savings target may provide a clearer incentive structure for the program administrator, and there is generally less controversy over whether the target is achieved. The fact that incentives are usually based on a calculation of shared benefits, where the predominant share of benefits goes to ratepayers, creates an equitable incentive structure: the program administrator receives fewer benefits and even if attributed (net) savings are lower than expected, the ratepayers still receive most of the benefits. For example, under an 80%–20% split of the benefits (80% of benefits are realized by ratepayers and 20% by the administrator), having attributed savings reduced by 50% still implies that 70% of the benefits go to ratepayers. See Rufo (2009) for other views on aligning incentives with the outputs of program evaluation.

[11] Symmetry refers to recognizing all the components of net savings – both positive and negative influences, and recognizing the impact on the net savings estimate when not all components are included.

[12] Agnew and Goldberg (2017) provide a number of choices for selecting comparison groups for use in billing analyses. This chapter also discusses using regression analysis as a tool for making appropriate comparisons and arriving at alternative net savings values.

[13] In this case, the nonparticipant baseline does not fully correct for free-riders, resulting in estimated net savings that are biased upward. Other self-selection factors could cause the participant and nonparticipant groups to behave differently. For example, if participants need the financial assistance to make the investment and nonparticipants do not need the rebate to take EE actions, the baseline comparison group might take more EE actions than the participant group in the absence of the program. In this case, a nonparticipant baseline would produce estimated net

Free-ridership reduces net program savings in this example case, but other variants of self-selection might increase net savings when a participant group is compared to a nonparticipant baseline. For example, if the customers who self-select into the program need the financial incentives to justify the EE investment, an adjustment for self-selection might increase overall net savings.

Spillover can also be viewed as a baseline issue. For example, nonparticipant spillover can occur when the energy consumption of the comparison group of nonparticipants is not indicative of what the energy consumption for this group would have been in the absence of the program. In this case, the comparison group is *contaminated:* the program affected the behavior of those in the comparison group.

This section discussed issues related to establishing an appropriate baseline as an approximation of the counterfactual scenario. Understanding that free-ridership, spillover, and market effects can be viewed as baseline issues can help the evaluator focus on the factors that are most important to selecting an appropriate method.[14] In many applications, selecting the baseline is a core issue in choosing an appropriate estimation method. When presenting the net savings results of a program, the evaluator should include a description of the baseline and the assumptions implicit in the estimation method.

---

savings that are biased downward and appropriately correcting for this self-selection effect would increase the estimated net savings. The authors have observed that often there is an assumption that addressing self-selection will always lower estimated net savings by reducing bias caused by free-riders, but this is not always the case.

[14] Self-selection, free-ridership, and spillover issues are not unique to EE evaluation—they are common in other settings as well. Consider a business decision made to produce net benefits, such as downsizing. Might self-selection be important to address in assessing this business initiative? Employees who have the best experience and are the most confident in their ability to land new jobs might (if able) self-select into the downsizing option. Might there be some free-riders if the downsizing effort includes personnel who were planning to leave anyway? Also, there might be spillover impacts from the downsizing program where having workers leave reduces the productivity of employees who remain. Although self-selection, free-ridership, and spillover pose challenges for EE evaluation, these same issues often have to be addressed in evaluating investment decisions in other fields and contexts.

4 Methods for Net Savings Estimation

This section discusses methods for estimating net savings, as well as some of the advantages and challenges associated with each. Evaluators use a variety of methods, some of which address free-ridership and/or spillover (for example, self-report surveys); others focus on market effects (for example, structured judgment approaches or historical tracing). The methods addressed in this section are:

- Randomized control trials (RCTs) and options for randomized approaches
- Quasi-experimental designs including matching
- Survey-based approaches
- Market sales data analyses
- Structured expert judgment approaches
- Deemed or stipulated NTG ratios
- Historical tracing (or case study) method
- Common practice baseline approaches
- Top-down evaluations (or macroeconomic models).

Table 1 lists methods that are applicable for estimating free-ridership, spillover, and market effects. This table indicates the general applicability of the methods. The following sections review the specific applications, caveats, limitations, and other key information in greater detail to explain how to assess the methods for each net savings component.

Table 1. Applicability of Approaches for Estimating Net Savings Factors

Method	Free-Ridership	Spillover	Market Effects
Randomized control trials and options for randomized approaches	Controls for free-riders[a]	Controls for participant spillover[b]	Not generally used
Quasi-experimental designs including matching	Controls for free-riders[c]	Controls for participant spillover	Not generally used; however, if the design includes observations over multiple years, then some market effects can be captured
Survey-based approaches	Is applicable	Is applicable	In conjunction with structured expert judgment
Market sales data analysis	Is applicable	Is applicable	Is applicable
Structured expert judgment[d]	Is applicable	Is applicable	Is applicable
Deemed or stipulated NTG ratios	Is applicable	Is applicable	Not generally used
Historical tracing	Is applicable	Is applicable	Is applicable
Common practice baseline methods	Is applicable	Not applicable[e]	Not applicable
Top-down evaluations	Assess the overall attributable change in energy use, so no separate adjustment is needed for free-ridership, spillover, and market effects[f]		

[a] Does not provide a direct estimate of free-ridership, but rather controls for free-riders through experimental design.
[b] Does not estimate spillover, but rather controls for participant spillover through experimental design. A separate study of control group members is required to address nonparticipant spillover if it is expected to be significant and affect the net impacts.
[c] Like RCTs, these designs do not provide a direct estimate of free-ridership, but self-selection bias can still be present. Unlike RCTs, the choice aspect of opt-in EE programs may not be fully addressed, unless additional methods are applied.
[d] This approach is applicable only if the experts are knowledgeable about the specific market being studied.
[e] Spillover could arguably be addressed through surveys of participants and nonparticipants, but this is not generally viewed as being part of the common practice baseline method, and the use of surveys would make this more similar to survey-based estimation methods discussed in Section 4.3.
[f] However, depending on the details of the analysis, these elements may not be fully captured.

More discussion on applicable methods for different types of residential and commercial programs and the pros and cons of these different methods can be found in a 2014 supplemental guidance document prepared for NEEP (NEEP 2014).

4.1 Randomized Controlled Trials and Options for Related Randomized Approaches

This section discusses random controlled trials (RCTs) and options for related random assignment approaches. RCTs represent the ideal approach and produce net savings accounting for free-ridership, participant spillover and avoid the problem of self-selection by addressing the potential choice-based biases by random assignment. However, RCT approaches may not always be possible. When an RCT is not possible, the quasi-experimental designs in Section 4.2 can be used as alternatives. RCTs can be difficult to set up and more applications are seen with pilot programs. However, RCTs are increasingly being used to evaluate behavioral programs, information programs, and pricing programs designed to increase efficiency. Generally, most RCT applications have been in the residential sector where large numbers of customers (both participants and controls) are available to the researcher. Even if a pure RCT is not possible, other approaches can be used to take advantage of random assignment. These other approaches, including random encouragement designs (RED) and random recruit or deny or structural criteria to avoid opt-in biases such regression discontinuity designs (RDD), have seen their applications to EE evaluation increase.

4.1.1 Randomized Control Trials (RCT)

An RCT design is ideal for assessing the net impacts of a program—particularly the free-ridership and short-term spillover components. If the RCT is short term (that is, 1 year or less), it may not be able to address any longer-term spillover, and addressing spillover and market effects may require additional data collection efforts for each year of the study.

For the RCT, the study population is defined first, then consumers from the study population are randomly assigned to either a treatment group (participants in the EE program) or to a control group that does not receive the treatment (nonparticipants). Random assignment is a key feature of this method. By using random probability to assign consumers to either the treatment or the control group, the influence of observable differences between the two groups is eliminated (for example, location of home, age of home, and appliance stock). Unobservable differences are also eliminated (for example, attitudes toward energy use, expectations about future energy prices, and expertise of household members in areas that might induce participation) (NMR Group, Inc. and Research Into Action [2010]; SEE Action [2012a, 2012b]). This method, when implemented properly, can provide a near-perfect baseline that results in reliable net savings estimates addressing free-riders and self-selection.

The net savings calculations are relatively straightforward when an RCT is designed properly. The literature generally covers three methods for calculating net savings:

1. **Use a simple post-period comparison to determine the differences in energy use between the control and treatment groups after participation in the program.** For example, if participating households are using 15,000 kilowatt hours (kWh) on average and the control households are using 17,000 kWh, the net savings estimate is 2,000 kWh.
2. **Use a difference-in-differences (DiD) approach to compare the change in energy use for the two groups between the pre- and post-participation periods**. For example, assume participants used 17,500 kWh prior to program participation and 15,000 after participation, for a difference of 2,500 kWh between the pre- and post-periods. Assume also that the well-matched control group has similar pre-period energy use (approximately 17,500 kWh), but the group's post-period energy use is 17,000 kWh (that is, slightly lower, possibly because of weather), for a difference of 500 kWh. Applying the DiD method results in an estimated savings of 2,000 kWh (the 2,500 kWh change for participants minus the 500 kWh change for nonparticipants).
3. **Use a linear fixed-effects regression (LFER) approach**, where the regression model identifies the effects of the program by comparing pre- and post-program billing data for the treatment group to the billing data for the control group. A key feature of the LFER

approach is the addition of a customer-specific intercept term that captures customer-specific effects on electricity use that do not change over time, including those that are unobservable. Examples include the square footage of a residence, the number of occupants, and thermostat settings (see Provencher and Glinsmann [2013] for an example and additional discussion of the LFER method).[15]

Even if randomizing the treatment and control groups, an evaluator may use a method other than the simple post-period comparison to be as thorough as possible and use all the available data to develop the estimate. The DiD method tracks trends over time, and the fixed-effects component of the LFER adds an extra control for the differences between consumers that are constant during the period being examined. All three methods generate unbiased estimates, as randomization ensures no systematic differences between the treatment and control groups in the drivers of energy use, so the three methods would be expected to generate similar, but not necessarily identical, results.

The RCT approach is simple in concept, but may be more difficult to implement given available data, timing, program design, and program implementation issues. It is becoming standard practice for evaluators to use statistical methods to test whether the allocation of customers between the treatment group and the control group is consistent with what would be expected from a random assignment of consumers to the treatment and control groups.[16] For billing data, this type of analysis often involves comparing the means of the two groups with respect to demographic variables (if available) and monthly energy use in the pre-program year. For example, if the differences in means for the two groups fall outside a 90% confidence bound for more than 2 months of the pre-program year, there is cause for concern that assignment to the two groups is not random. (See an example of an application of this test for consistency with RCT expectations in Provencher and Glinsmann [2013] and other tests in Stuart [2010]). If this is the case, it is worth examining how the random assignment was conducted to ensure no inadvertent elements of the process are affecting assignment to the treatment and control groups. The goal of this testing is to determine if non-random factors are affecting the assignments, not to keep repeating the random selection process until the samples fit an ideal profile. If several characteristics are compared, it is not unusual to have some that are "significantly" different between the two groups. Regression analysis helps to mitigate these effects.

The RCT approach to estimating program impacts reflects the "intent to treat" effect. Generally, it is not appropriate to drop customers after the random assignment, though the consequences of doing so vary. For example, questions may arise about what to do with consumers who opt out.

[15] A number of the methods discussed in this chapter use regression approaches. Some are fairly simplistic; others are quite sophisticated, requiring expertise in econometrics. Each section provides citations to applied studies, many of which describe the econometric techniques employed. For example, Stuart (2010) lists econometric software and routines that can be useful in matching. Also, Agnew, and Goldberg (2017) discuss regression models in more detail but provide a limited set of literature references. SEE Action (2012a) recommends Greene (2011) as a useful reference on regression techniques. Wooldridge (2010) focuses on cross-section and panel data models that are often used in evaluation. Kennedy (2008) and Angrist and Pischke (2008) are useful supplements to any econometrics textbook.

[16] Even with random assignment, it is important to apply best practices in the design of analysis including stratification both to reduce standard errors (increase precision) and help ensure representativeness of the sample drawn.

Consider, for instance, a program involving Home Energy Reports (HERs), in which program administrators send energy use reports by mail. This program was designed to generate energy savings by providing residential consumers information about their energy use and energy conservation. Some percentage of consumers will opt out of the program. They should remain in the analysis because the similar set of control consumers who would have opted out of the program could not be identified if they were to receive the report. Also, on average, these consumers might have different energy use than the other control consumers, causing the reported impact to be biased if the treatment group is adjusted to remove the opt-out consumers. At the other extreme, HERs might not be deliverable because of observable address characteristics. If this same address characteristic can also be identified for control consumers, the estimate of program impacts after eliminating treatment and control consumers with this characteristic is, strictly speaking, an unbiased estimate of the effect of intent-to-treat *conditional* on the address characteristic. These examples are meant to show that careful analysis is needed in the application of all methods, including RCTs. In addition, Duflo et al. (2007) caution that excessive investigations of subgroups not specified *ex ante* constitute a form of data mining that should be avoided. The case discussed above where address characteristics are available for the treatment and control groups does not fall in this category, but this caution deserves emphasis.

To maintain an RCT over a period, evaluators must take care when working with the data across the treatment and control groups. For example, a behavioral program (such as HERs) may be rolled out to 20,000 high-use residential consumers in program year 1. In program year 2, an additional 20,000 consumers of all energy use classifications may enroll, and another 30,000 consumers may enroll in program year 3. Additionally, some consumers in program year 1 may have dropped out (requested to not receive the HERs).[17] Each of these sets of participating customers need to be appropriately considered in the RCT design and the appropriate assignment of customers to be used as controls.

Issues inevitably arise about the consumer energy use data. Researchers have used the following criteria, among others, as indicators of problems with consumer billing data:

- Having fewer than 11 or more than 13 bills during a program year
- Having fewer than 11 or more than 13 bills during the pre-program year
- Energy consumption outside a reasonable range (that is, an outlier observation with average daily consumption that is lower than the 1st percentile or higher than the 99th percentile)
- Observations with fewer than 20 or more than 40 days in the billing cycle.

Agnew and Goldberg (2017) also discuss issues with consumer energy use data and program data in residential settings. Even programs that have operated for several years are likely to have issues. Using the HERs example, this could include consumer records that are missing the date when the first report was sent or entries in consumer records that indicate issues with that observation.

[17] This is not an unusual problem in the utility industry. Utilities have for many years addressed similar issues in maintaining random customer samples for load research.

After addressing data issues, the evaluator probably still has a good RCT, unless many consumers are affected by these data issues or consumers are disproportionately affected across the participant and control groups. Mort (2017) presents additional criteria that can cause sites to be excluded and suggestions about what to do if the number of removed sites exceeds 5%.

The ability to disseminate information to large groups of consumers has led to an increase in RCTs in EE evaluation.[18] In general, these RCT-based evaluations have focused on residential behavior-based EE programs such as HERs programs. These programs lend themselves to random trials in that they: (1) provide information only; (2) can be implemented for relatively homogeneous consumers at the same time; and (3) allow for an RCT design. These characteristics, however, are not generally present for many large-scale EE programs that tend to account for many of the EE portfolio savings.

In summary, the RCT approach is the most accurate method for estimating net impacts. The RCT controls for free-riders and near-term participant spillover—two important factors. To the extent that the program affects the control group, nonparticipant spillover is not addressed. This effect is likely to be small over the short run in most behavioral programs. If nonparticipant spillover is large, net impacts will be underestimated because there are nonparticipants who were affected by the program, and the baseline will be inaccurate. To appropriately address this issue, the evaluator would need to conduct a separate study of control group members to address nonparticipant spillover. Because market effects are longer term spillover effects, they would likely not be included in any RCT net savings approach that spans just a few years.

Although the RCT method can produce an accurate baseline when constructed correctly, it is not always possible to apply an RCT to evaluations of EE programs for a variety of reasons. RCT generally requires planning in advance of program implementation. As pointed out in Chapter 8 (Agnew and Goldberg 2017) of these protocols, "…evaluation concerns have been less likely to drive program planning." Also, an RCT approach may involve denying or delaying participation for a subset of the eligible and willing population. In some cases, the random assignment may result in providing services to consumers who either do not want them or may not use them (see Table 2 for pros and cons of RCTs).

Other characteristics of programs that can make an RCT difficult to implement include:

- Programs that require significant investments, such as a commercial and industrial (C&I) major retrofit program in which the expenditures are in the tens of thousands of dollars. Typically, these programs are opt-in, and random assignments within an eligible study population might include consumers who either do not need the equipment or services or do not want to make that investment. Programs that involve relatively large investments

[18] Evaluations of HERs programs that used RCTs include Sacramento Municipal Utility District (2011), Puget Sound Energy (2012), AEP (2012), PG&E (2013), Commonwealth Edison (2012), and Pacific Gas & Electric (2013). Some ongoing evaluations use RCT methods for HERs programs, and will produce additional practical information on RCT applications. Another useful study, but one focused on evaluating pricing programs, which used an RCT design is the Sacramento Municipal Utility District (2013). This study assesses different pricing structures in the residential sector; however, the methods used are good examples of what can also be applied in EE evaluations in an RCT context.

in measures and services across the residential and C&I sectors may not amenable to an RCT design.

- C&I programs often have participants that are more heterogeneous than is the case for residential programs which would require large samples of both treatment and control groups than may be available. In some cases, a few very large customers can be relatively unique within a utility service area or region with few similar consumers who might be appropriate candidates for a control group.
- To achieve savings targets, programs may be rolled out over an entire year, with consumers opting in every month. As a result, consumers self-select into the participant group, which is unknown until after 1 year of the program implementation. Evaluators can more easily apply RCT to programs with a common start date for many participants (for example, HERs programs). There are ways to address this, but this adds somewhat to the complexity of the design. The random recruit and deny design discussed below can be used to addressing rolling program roll-outs.

4.1.2 Other Forms of Randomization and Approaches for Minimizing Opt-in Selection Bias

Two other approaches incorporate random assignment to help address the choice-based bias of opt-in programs—random encouragement design (RED) and random recruit deny/delay approaches. Another approach that can be used to minimize opt-in selection bias is the regression discontinuity method (RDD). RDD does not incorporate randomization. Instead, it looks for cutoffs or discontinuities in participation that can be used to construct two eligible participant groups.

4.1.2.1 Random Encouragement Design

Random encouragement design (RED) is also applicable to the types of data available for EE program evaluation. Rather than being randomly assigned to a treatment or control group, customers are randomly assigned to receive supplemental encouragement to participate in the program (e.g., a letter informing a random set of customers about a rebate), or not to be so encouraged. RED involves taking a randomly selected group of participants to receive extra encouragement, which typically takes the form of additional information or incentives. A successful encouragement design allows the effects of the intervention and encouragement to be estimated (Diamond and Haninmueller 2007; McKinzie 2009[19]). In this case, there may be an EE program for which all consumers can decide to opt in such as a residential audit program or a commercial audit or controls programs. A group of randomly selected consumers is then provided extra encouragement in terms of information and/or financial incentives. This randomization can ameliorate the effects of self-selection.[20]

[19] In a position statement closely related to what EE program evaluators face, McKenzie states that "Rigorous impact evaluations, which compare the outcomes of a program or policy against an explicit counterfactual of what would have happened without the program or policy, are one of the most important tools that can be used along with appropriate economic theory for understanding 'what works.' Despite this, until recently impact evaluations have been rare, especially outside the areas of health and education."

[20] The underlying estimation concept in RED is explained by the U.S. Department of Energy (2010): "In RED, researchers indirectly manipulate program participation using an encouragement 'instrument' so as to generate the

The RED design provides the average net savings per participant for those who participate because of the encouragement but otherwise would not. This is not necessarily the same as the net savings for the original program without extra encouragement. In particular, we would expect free-ridership to be lower among those who need extra encouragement. Thus, the RED might be expected to overstate net savings for the original program if free-ridership is present but would still provide useful information.

Fowlie and Wolfram (2009) outline an application of RED to a residential weatherization program and address the design of the study and Fowlie, Greenstone and Wolfram (2015) apply this design to a low-income program.[21] They point out that:

REDs are particularly useful when:

- *Randomization of access or mandatory participation is not practical or desirable.*
- *There is no need to ration available services (that is, demand does not exceed supply).*
- *The effects of both participation and outreach are of interest to policy makers.*

Rather than randomize over the intervention, the encouragement to participate is randomly manipulated. This allows the effect of the encouragement to produce exogenous variation in program participation, which can help identify the effect of the program on participants (U.S. Department of Energy 2010).

Evaluators should take certain practical issues into account in any research design, and RED is no exception. The sample sizes needed for an RED study are typically larger than for a pure RCT, and groups receiving the encouragement need to show different participation rates.[22] Evaluators should consider this research design when estimating net savings, as it aligns well with many standard EE program implementation plans. The random variation is designed not by excluding participants but simply by providing enhanced information and/or incentives offered to the selected consumers. Ongoing research work using RED should provide useful information for practitioners and the EE evaluation community. RED is growing with most applications focused on residential programs.

4.1.2.2 Random Recruit Deny/Delay

Finally, another approach that can be used to construct randomized treatment and control groups is the random recruit and deny/delay design. In this case, the timing of the treatment is randomly assigned. Customers are recruited with the understanding that they will randomly be assigned to receive the program offering immediately or later. The control group is thus a randomly selected set of customers that have opted-in but receive the treatment later. This is an effective way of

---

exogenous variation in program participation that is so essential for causal inference. This exogenous variation can then be used to identify the effect of the program on those households whose participation was contingent upon the encouragement." Other useful references to RED are Bradlow (1998) and West (2008).

[21] Fowlie et al. (2015) find limited energy savings from the weatherization assistance program that was evaluated. This find was challenged by Hogan (2015).

[22] This can be one of the challenges in the design of an RED approach. The design of the encouragement given to a random sample of participants must be effective; that is, produce higher acceptance rates than for the balance of the participant group.

ensuring that the control and treatment groups are well matched, but it may not fully address other types of selection. If it is a pilot program that is being evaluated, the savings impacts may be accurate for the pilot participants, but it may be difficult to extrapolate these impacts to a broader set of customers. There are two other issues: (1) The fact that a customer did opt-in to a program but had their participation delayed may, in itself, change their behavior (e.g., they may not take actions they would have taken in the absence of the program as they are expecting to receive the benefits of participation in the near future); and, (2) Some customers may drop out of the research pool if they learn their participation will be delayed and that they will be part of a control group. An example of this research design is Xcel Energy (2016).

4.1.2.3 Regression Discontinuity Design

SEE Action evaluation guides (2012a, 2012b) discuss the regression discontinuity design (RDD). This method is becoming more widely used and is applicable to programs where a cutoff point or other discontinuity separates otherwise likely program participants into two groups. This approach examines the impacts of a program by using a cutoff value that puts consumers into or out of the program through a design that does not involve their selecting themselves into the program or choosing not to participate. As a result, this approach addresses the self-selection issue.[23] By comparing observations lying closely on either side of a cutoff or threshold (i.e., the eligible and in-eligible cut off), the average treatment effect in environments where randomization is not possible can be estimated.[24] The underlying assumption in RDD is that assignment to participant and nonparticipant groups based on the eligibility cutoff produces groups that are otherwise similar. If this holds, those who just met the threshold for participating are comparable to those who just missed the cutoff and did not participate in the program and the difference in energy use between the two groups can reasonably be assumed to be the effect of program participation.

The SEE Action reports indicate that RDD can be a good candidate for yielding unbiased estimates of energy savings. The example used by SEE Action is based on an eligibility requirement for households to participate in a program. This requirement might be that a consumer whose energy consumption exceeds 900 kWh/month would be eligible to participate in a behavior-based efficiency program, while consumers who use less than 900 kWh/month would be ineligible. Thus, the group of households immediately below the usage cutoff level might be used as the comparison group.

For participating and nonparticipating households near the cutoff point of 900 kWh in monthly consumption, RDD is likely to be a good design. In the larger context, this RDD assumes that the program impact is constant across all ranges of the eligibility requirement variable (that is, the impact is the same for households at all levels of energy use). Evaluators should consider this

[23] In the recent years, there has been a strong movement toward focusing on the “identification” issue in evaluation; that is, the issue that in the absence of an RCT you do not really know if the error term in a regression is correlated with the explanatory variable of interest, so your estimate of the coefficient on that explanatory variable should be assumed to be biased in the absence of “sound” corrective action. A regression discontinuity design addresses this issue.

[24] The RDD has a history in evaluation dating back to the 1960s. This approach has been used to assess a wide variety of attribution analyses in the fields of education, health, and policy. Recently, this approach has been used more often. For a review of RDD see Imbens and Lemieux (2010).

assumption carefully for participating households that might consume much more than 900 kWh/month (for example, 2,000 kWh or more for some participants). Households with greater consumption may have greater opportunities for energy use reductions (although the change might be constant as a percentage). In this example, potential concerns about the consistency of program impacts across different levels of household energy use suggests an assessment of the quality of the resulting participant and control groups matched samples. Stuart (2010) has general guidance for assessing the quality of these designs.

Another discontinuity example is a time-based cutoff point. Because utilities often have annual budgets for certain programs, it is not uncommon for a program to exhaust its budget before the year is finished, sometimes within 6 months. In this case, a date-based cutoff is useful. Consumers who apply for the program after the enrollment cutoff date imposed by budget restrictions may be similar to the program participants accepted into the program during the first 6 months of the year. Also, both groups of consumers may have a more similar distribution of energy use per month (the focus of an impact assessment). This time-based cut-off approach is similar to using future participants as comparison groups discussed in UMP *Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol* (Agnew and Goldberg 2017).

4.1.3 Summary – RCTs and Related Randomization Approaches

Several types of approaches employing randomization are being used in evaluation. These include RCTs, REDs, and Recruit Deny/Delay. The RDD approach is a bit different as it takes advantage of a structural discontinuity (participation cut-off or threshold) to a treatment and control groups that are not affected by choice (i.e., opt-in programs).

The RCT approach is the most accurate method for estimating net impacts. The RCT controls for free-riders and near-term participant spillover—two important factors. To the extent that the program affects the control group, nonparticipant spillover is not addressed. This effect is likely to be small over the short run in most behavioral programs. If nonparticipant spillover is large, net impacts will be underestimated because there are nonparticipants who were affected by the program, and the baseline will be inaccurate. To appropriately address this issue, the evaluator would need to conduct a separate study of control group members to address nonparticipant spillover. Because market effects are longer term spillover effects, they would likely not be included in any RCT net savings approach that spans just a few years. These same caveats also apply to RED, Recruit Deny/Delay, and RDD approaches.

It is not possible to definitively determine whether the RED or Recruit Deny/Delay designs discussed above provide an appropriate comparison group. Fowlie and Wolfram (2009) point out that there have been studies comparing these designs to the ideal RCT. The finding is that randomized designs (either RED or RDD) improve on simple comparison approaches. RDD depends on the program having a cutoff point for participation that allows for random selection. RED may be a good fit with many EE programs that have many participants, but appropriate design in the types of information and incentives is required. Both RDD and RED depend on the assumption that the net savings of the isolated participants—those just under the threshold for RDD, and those who participate with only incremental encouragement for RED—is the same as the net savings for all participants.

Importantly, these methods should be considered in advance of program implementation to allow for the appropriate data, or the design of the information or incentives that will be offered to potential participants, to effectively implement these evaluation methods. It has always been important to consider evaluation when designing or revising EE programs, but for random assignment methods the evaluation method must be built into the program delivery.

Some of the pros and cons associated with these methods are presented in Table 2.

Table 2. Approaches using Random Assignment (RCTs, RED, Recruit Delay/Deny, and RDD) — Summary View of Pros and Cons

Pros	Random assignment reduces and limits bias in estimates Increases reliability and validity RCTs control for free-riders and participant spillover RCTs widely accepted in natural and social sciences as the highest standard of research designs RED, RDD, and Recruit Delay/Deny approaches also control for free-riders and participant spillover, but with additional assumptions regarding the appropriateness of each design.
Cons	Bias can result if random assignment occurs among volunteers or if the program drop-out rate differs by key characteristics Does not address nonparticipant spillover Equity/ethical concerns about assigning some ratepayers to a control group and not allowing them to participate in the program for a period of time May not be applicable to programs that involve large investments in measures and services Some C&I programs can have participants that are unique due to their size or industry, and there may be few control group candidates Needs to be planned as part of program implementation to allow for appropriate random assignment for RCT, RED, and Recruit Deny/Delay.

**This summary of pros and cons is not meant to replace the more detailed discussion in the text for guidance in application*.

4.2 Quasi-Experimental Designs

For most EE programs, either practical concerns or design factors will limit the use of RCT and other random assignment methods. In these situations, quasi-experimental designs are often a good option. Quasi-experimental designs are not unique to EE evaluations and are often used in evaluations of private and public investments. Stuart (2010) reviews the evolving research on matching and propensity scoring methods in quasi-experimental designs and states that such methods "… are gaining popularity in fields such as economics, epidemiology, medicine, and political science."[25,26]

[25] Stuart (2010) also provides a guide to software for matching, because software limitations have made it difficult to implement many of the more advanced matching methods. However, recent advances have made these methods more accessible. This section lists some of the major matching procedures available. A continuously updated version is also available at www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html. Common statistical software packages such as STATA, SAS, and R address most of the current matching approaches.

Quasi-experimental designs have some similarities to RCTs in terms of constructing comparison and treatment groups, except that random assignment is not possible. In a quasi-experimental design, consumers typically select themselves into the participant group, and the evaluation researcher must then develop the comparison group. To avoid confusion, quasi-experimental designs use the term *comparison group*, and RCT designs use the term *control group*. [27]

This section discusses two types of approaches to developing a comparison group within a quasi-experimental design–1) matching methods, and 2) panel data approaches. Matching methods use a measure of distance between two observations (e.g., customers) and can include Exact Distance, Mahalanobis Distance, and Propensity Scoring. Panel data approaches include structural regression modeling with a specific set of independent variables designed to address differences between the treatment and comparison groups (see the discussion of pooled regression in Agnew & Goldberg, 2017).

4.2.1 Matching Methods

Matching is broadly defined in the literature to be any method that aims to equate (or balance) the distribution of covariates in the treatment group and the comparison group.

The evaluator's goal is to select a comparison group that matches the participant group in terms of the actions that influence energy use. If done well, the only significant difference between the two groups will be participation in the program. Still, how well the comparison group actually matches the participant group will always be subject to some uncertainty, as there may be *unobservable* variables that affect energy use, the attribute of interest. Stuart (2010) defines the problem this way:

> *One of the key benefits of randomized experiments for estimating causal effects is that the treated and control groups are guaranteed to be only randomly different from one another on all background covariates, both observed and unobserved. Work on matching methods has examined how to replicate this as much as possible for observed covariates with observational (nonrandomized) data… While extensive time and effort [are] put into the careful design of randomized experiments, relatively little effort is put into the corresponding "design" of nonexperimental [quasi-experimental] studies. In fact, precisely because nonexperimental studies do not have the benefit of randomization, they require even more careful design.*

---

[26] Most attribution analyses assessing business decisions and public or private investments use quasi-experimental designs, as many practical factors result in the use of this method. As an extreme example, consider a study that is designed to assess the health effects of smoking. Would it be appropriate to select a study population of 9,000 18-year-olds and assign one third to a group that does not smoke, one third to a group that smokes a pack of cigarettes a day, and one third to a group that smokes a pack a day, but with some mitigating medications? Clearly, this type of RCT would pose ethical issues. As a result, natural quasi-experiments are used where smokers are matched with a comparison group of nonsmokers that is as representative as possible. The methods of matching on observable characteristics have become quite advanced in the past decade.

[27] Technically, quasi-experimental designs do not always include a nonparticipant comparison group. For example, the interrupted time-series design (Shadish et al. 2002) relies only on aggregate participant data over time and shows this method can help control for threats to internal validity; i.e., that the results of the study are appropriately estimated for the participating customers. External validity involves generalizing; i.e., the ability of the study results to be extrapolated to other groups of customers.

Stuart (2010) presents a good overview of the literature on matching and advantages of matching compared to regression models based on a set of explanatory variables. The recent evaluation literature, particularly for residential sector programs, shows the increasing use of matching. Recent approaches have focused on matching by energy use and energy use distributions across months and seasons. These matching methods can be simple or sophisticated, even when matching is confined to available energy use data (that is, no additional surveys of nonparticipants are conducted). Matching on energy use can be as simple as stratifying participants and nonparticipants by their energy consumption (season, year, or month) and then drawing nonparticipants to match the participants' distribution of energy use.

As discussed by Stuart (2010), the literature on matching based on energy use is expanding. Provencher and Glinsmann (2013) focus on a comparison of the distribution of energy across months and seasons. The analysis follows the approach advocated by Ho et al. (2007) and Stuart (2010). The procedure used by Provencher and Glinsmann (2013) first matches each participant household to a comparison household based on a minimum distance criterion—in this case, the minimum sum of squared deviations in monthly energy consumption for the 3 months of the specified season in the pre-program year.[28] In the second step, a regression model of the energy use of treatment customers and their matched controls, with covariates that include the matching variables, is used to identify the average treatment effect.

Matching methods tend to follow the literature reviewed by Stuart (2010). Stuart indicates that matching methods have four key steps, with the first three representing the "design" and the fourth the "analysis." These steps are:

1. Define closeness: the distance measure used to determine whether an individual is a good match for another.
2. Implement a matching method appropriate to the measure of closeness.
3. Assess the quality of the resulting matched samples (and perhaps iterate Step 1 and Step 2 until well-matched samples result).
4. Analyze the outcome and estimate the treatment effect, given the matching done in Step 3.

In Step 1, closeness is often defined as a minimum distance value as used in Provencher and Glinsmann. Another approach for identifying nonparticipants is "propensity scoring." The most common method used in propensity score estimation involves the estimation of a logistic regression. This model uses information about participants and nonparticipants to estimate a dependent variable assigned the value of 1 if that consumer is a participant or 0 if the consumer is a nonparticipant. This process allows for identification of nonparticipants who have similar

[28] In the program evaluation literature, matching often involves matching on variables with different metrics; for example, energy use and square footage of the household. These variables are normalized in the application of the distance criterion, usually using the full covariance matrix for the variables (the Mahalanobis metric). The original reference is Mahalanobis (1936) and the use of the metric is covered by Stuart (2010). One application, among many examples, is Feng (2006), which also includes the SAS® code for this method.

propensity scores to participants (that is, similar attributes between participants and nonparticipants). This approach has a long history in in the EE evaluation literature.[29,30]

The EE evaluation literature using matching methods (i.e., approaches that use a definition of closeness) has been expanding. Different types of applications that develop matching subject to constraints (e.g., a geographic constraint such as falling within a defined set of zip codes – See Navigant, 2016), and matching on hourly consumption rather than monthly data (PowerStream, 2016) are becoming more common. An application to a commercial sector pricing/thermostat program is found in Nexant (2017).

4.2.2 Panel-Data Models

Stuart (2010) states that alternatives to matching methods include adjusting for relevant covariates in a structured regression model. However, Stuart (2010) also points out that "matching methods should not be seen in conflict with regression adjustment and in fact the two methods are complementary and best used in combination."

One of the motivations for matching is to mitigate against model specification bias in the traditional structured regression panel-data model. Chapter 8 of the UMP (Agnew and Goldberg 2017) discusses consumption data analyses, including alternatives for constructing comparison groups. Also, the two SEE Action guides (2012a and 2012b) address matching. Matching methods include:

- **Participants as the comparison group:** SEE Action (2012b) states that among quasi-experimental approaches, "perhaps the most common [is] the 'pre-post' approach. With this approach, sites in the treatment group after they were enrolled in the program are compared with the same sites' historical energy use prior to program enrollment. In effect, this means that each site in the treatment group is its own nonrandom control group."

  By using the participant group as its own comparison group, the energy use of the participants during a period before they participated in the program is used as the

---

[29] The use of discrete choice methods to address self-selection in evaluations of EE programs has been presented in early evaluation handbooks. See Violette et al. (1991) and Oak Ridge National Laboratory (1991). More recently, Bodmann (2013) used a discrete choice model to develop an instrumental variable to address omitted variable bias. However, most of these applications occurred in the 1990s, probably because the development of a discrete choice model that has adequate predictive power requires large sample sizes, which make the surveys expensive to conduct. The discrete choice model needs to be able to predict customers who choose to participate and customers who choose not to participate with appropriate reliability. This approach thus requires both participant and nonparticipant surveys. This more advanced econometric topic is not dealt with in detail in this chapter; however, several reviewers believed it was important to provide references to these methods. Heckman (1979) originally developed the two-stage model for treating self-selection. These techniques are addressed both under instrumental variables and self-selection by Kennedy (2008), who states: "Selection is not well understood by practitioners. It rests fundamentally on the role of an unmeasured variable and so is similar to bias created by the omission of a relevant explanatory variable." (p. 286). An updated discussion of the Heckman models for self-selection, along with appropriate caveats, can be found in Guo and Fraser (2010). Note: a link to this chapter is provided in the References section. Guo and Fraser also show how the Heckman models relate to propensity scoring. Applications in the EE arena include Dubin and McFadden (1984), Goldberg and Kademan (1995), and Bodmann (2013), who used a discrete choice model to develop an instrumental variable to address omitted variable bias.

[30] Southern California Edison (2014) provides a recent behavioral impact application using propensity scoring.

comparison or baseline. A statistical consumption analysis is used that also includes factors that are expected to influence energy use and may vary across the pre-post time periods. Weather is the most obvious additional variable that should be controlled, but there may be other variables as well, such as economic factors if the periods cover a two-year period or longer. Agnew and Goldberg (2017) provide a useful set of algorithms for making weather adjustments.[31]

- **Nonparticipants as the comparison group:** The trend in the literature is to move away from the simple approach of using participants as their own comparison group in a time-series analysis and instead to develop cross-sectional time-series data that include data on participants and nonparticipants.

4.2.3 Summary of Quasi-Experimental Designs—Matching and Panel Data Regression Models

Randomized approaches may not always be possible to use. Quasi-experimental designs try to replicate designs that employ randomization using observational (nonrandomized) data. Matching as an evaluation method is rapidly expanding, particularly for residential programs. Panel data regression models can be used in conjunction with matching, or they can be used as stand-alone methods when data are available on relevant covariates and there is confidence in the appropriate structure for the models. Table 3lists some pros and cons with these approaches.

Table 3. Quasi-Experimental Designs—Summary View of Pros and Cons

Pros	Limits bias if a matched comparison group can be identified regarding the actions that influence energy use Unlike RCT, can be applied after program implementation Increases reliability and validity Partially controls for free-riders and participant spillover Widely accepted in natural and social sciences when random assignment cannot be used Matching may reduce concerns over model specification bias.
Cons	May be difficult to identify a matched comparison group if there are unobservable variables that affect energy use Does not address nonparticipant spillover Some C&I programs may have unique participants and few control group candidates Does not address self-selection bias without additional modeling, i.e., the estimation of a companion discrete choice participation model to address bias from choice-based participation in programs.

4.3 Survey-Based Approaches

This section describes the survey-based approach to collect NTG-related data and the analytic use of the data obtained. This approach can be a cost-effective, transparent, and flexible method for estimating NTG, and it has become one of the most often-used methods in EE net savings

[31]Other approaches can be used for weather normalization, particularly if the evaluator is interested in changes in monthly peak demand in addition to average monthly energy use. Additional weather normalization approaches are discussed by Eto (1988) and McMenamin (2008).

estimation. Consequently, it is important to understand good sample and survey design, and the strengths and weakness of these methods.

Surveys may target up to three types of respondents: (1) program participants, (2) program nonparticipants, and (3) market actors.[32] This section individually describes surveys with these three types of respondents; best practices recommend triangulating and using multiple survey approaches (for example, enhanced self-report) or multiple net savings estimation approaches.

The methods discussed in the preceding section provide estimates of net savings directly. That is, those approaches compare a participant group to either a random control group (as part of an RCT) or to a comparison group from a well-designed, quasi-experimental application, and these approaches do not require a separate effort to estimate free-ridership, spillover, or market effects.[33]

Survey-based approaches are used in evaluations that start with gross estimates, and then adjust for NTG factors. Surveys can be a cost-efficient means to estimate NTG factors, but they are not without issues, as discussed in the following subsections. Baumgartner (2013) also discusses many of the issues involved in using surveys to estimate NTG.

4.3.1 Program Participant Surveys

Survey-based methods for estimating net savings from program participants who are aware of the program incentives/services use questions about the program's influence on the participants' actions and decision-making. Participants answer a series of closed-ended and open-ended questions on these topics:

- Why they installed the program-eligible equipment.
- What they would have done in the absence of the program incentive and services.
- What further actions they took on their own because of their experiences with the program.

As noted by Baumgartner (2013), best practice survey design for attitudes and behavior measurement use multiple-item scales to better represent the construct. Because participant decision-making is complex, the survey should ask a carefully designed series of questions rather than a single question, as that could result in misleading findings. Refer to SEE Action (2012b), Megdal et al. (2009), Haeri and Khawaja (2012), and New York Department of Public Service (2013b) for discussions about the sequencing of a series of questions.

The primary benefits of a survey-based approach are:

- A survey approach can be less expensive than other approaches, particularly if the effort is combined with data collection activities that are already planned for process and impact evaluations.

[32] Note that a Delphi panel, which also uses surveys of a panel of experts, is discussed in Section 4.5 of this chapter.
[33] Market effects can be viewed as longer-term spillover effects; therefore, it is unlikely that any market effects are included in an RCT net savings approach spanning just a few years.

- The evaluator has the flexibility to tailor questions based on variations in program design or implementation methods.
- It can yield estimates of free-ridership and spillover without the need for a nonparticipant control group (NMR Group, Inc. and Research Into Action 2010). However, participant surveys capture only a subset of market effects,[34] a key piece of NTG.

Despite these benefits and the wide use of a survey-based self-report approach, significant concerns have been raised (Ridge et al. 2009; Peters and McRae 2008). The main concerns are:

- A potential bias related to respondents giving socially desirable answers.[35]
- The inability of consumers to know what they would have done in a hypothetical alternative situation, especially in current program designs that use multiple methods to influence behavior.
- The tendency of respondents to rationalize past decisions.
- A potential for arbitrariness in the scoring methods that translate responses into free-rider estimates.
- Consumers may fail to recognize the influence of the program on other parties who influenced their decisions. For example, a program having market effects may have influenced contractor practices, which in turn may have indirectly impacted the participants' (and nonparticipants') decisions.

Ridge et al. (2009) point out that, although these concerns are valid, they are widely acknowledged by social scientists who have worked on a variety of methods over the years to address them. It is also important to recognize that all methods have potential biases.[36] For example, market sales analysis,[37] which is based on objective sales data, can be biased if the market actors who provide data for the analysis operate differently from those not participating in the study or if the comparison area is systematically non-comparable.

In addition, Ridge et al. (2009) point out that it does not make sense to compare all self-report approaches equally, as some conform to best practice and others do not. Keating (2009) adds that many of the criticisms of the self-report approach can be alleviated through careful research design, sampling, survey timing, and wording of questions.

Baumgartner (2013) presents guidelines for selecting appropriate survey designs and recommends procedures for administering best practice surveys. The literature also contains

[34] Participant surveys can, in theory, capture end user market effects; for example, changes in end user awareness, knowledge, and efficiency-related procurement practices.
[35] Participants may also have a bias toward overstating program impacts because they want to retain incentives, although this has not been widely documented.
[36] This is, of course, the primary motivation for triangulation.
[37] Market sales analysis captures the total net effect of a program. Ideally, this method involves obtaining comprehensive pre- and post-market sales data in both the area of interest and in an appropriate comparison area and examining the change in the program area compared with the change in the non-program area (Tetra Tech et al. 2011).

several best practice elements for survey design, data collection, and analytic methods specific to estimating net savings (New York State Department of Public Service 2013; Tetra Tech et al. 2011). This literature notes the importance of making the entire process transparent so stakeholders can understand how each question and its response impacts the final estimate. Thus, the report should contain details of critical elements such as the question sequence, scoring algorithms, and the handling of inconsistent and/or missing data.

4.3.1.1 Survey Design Elements

Several design elements need to be considered when developing surveys. Best practices for choosing design elements include:

- Identify the key decision-maker(s) for the specific EE project. For downstream programs, a key decision-maker in the household or business is likely to be responsible for making the final decision, although they may assert that their vendor was the most influential in their decision. Although consumers ultimately decide what they will purchase, they may not be aware of the influence of the interventions for upstream programs where trade ally decisions are driving change (for example, original equipment manufacturers determine equipment EE levels and retailers determine what equipment to stock and market, or advertise as a result of upstream program incentives).
- Use setup or warmup questions to help the decision-maker(s) recall the sequence of past events and how these events affected their decision to adopt the measure.
- Use multiple questions to limit the potential for misunderstanding or the influence of individual anomalous responses.
- Use questions that rule out rival hypotheses for installing the efficient equipment.
- Test the questions for validity and reliability.
- Use consistency checks when conducting the survey to immediately clarify inconsistent responses.
- Use measure-specific questions to improve the respondent's ability to provide concrete answers, and recognize that respondents may have different motivations for installing different measures.
- Use questions that capture partial efficiency improvements (accounting for savings above baseline but less than program eligible), quantity purchased, and timing of the purchase (where applicable for a measure) to estimate partial free-ridership.
- Use neutral language that does not lead the respondent to an expected answer.
- Use combinations of open- and close-ended questions to balance hearing from the end users in their own words and create an efficient, structured, and internally consistent dataset.

4.3.1.2 Data Collection Elements

Even when the survey design is effective, data collection should also follow best practices for collecting reliable information and calculating valid estimates. These practices include:

- Pretest the survey instrument to ensure that questions are understandable, skip patterns are correct, and the interview flows smoothly. The pretesting should use, when possible, cognitive interviewing techniques (Miller 2011).[38]
- Use techniques to minimize nonresponse bias, such as advance letters on utility or program administrator letterhead (the organization for which the participant will most likely associate the program) and multiple follow-ups over a number of weeks.
- Follow professional standards for conducting surveys, which include training and monitoring interviewers.[39]
- Determine the necessary expertise of the interviewer based on the complexity and value of the interview (for example, it is better for trained evaluation professionals rather than general telephone surveyors to address the largest, most complex projects in custom programs).
- Time the data collection so it occurs as soon as possible after a measure is installed, as this minimizes recall bias and provides timely feedback on program design. Recognize, however, that timely data collection for estimating free-ridership will underestimate participant spillover, as little time may have passed since program participation. Conducting a separate spillover survey later with these same participants can alleviate this. Having a separate survey will increase data collection costs, but may be warranted if spillover effects are likely to have occurred.
- Sample (or oversample) a census of the largest savers and, depending on program participation, sample end uses with few installations to ensure the measures are sufficiently represented in the survey sample.

4.3.1.3 Analytic Elements

In addition to discussing survey design and data collection elements, much of the literature discusses best practices for analysis such as:

- Treat acceleration of the installation of the EE measures appropriately to produce lifetime net savings rather than first-year net savings (this requires understanding the program's influence on the timing of the project).[40]
- Incorporate the influence of previous participation in the program.

[38] In cognitive interviews, respondents are asked to describe how and why they answered the question as they did. Miller (2011) notes that "through the interviewing process, various types of question response problems that would not normally be identified in a traditional survey interview, such as interpretive errors and recall accuracy, are uncovered." (p. 54).

[39] Data collections surveys can be conducted via telephone, the Web (including smartphones), postal mail, and in person. For large complex C&I projects, an energy engineer who is knowledgeable about the type of project and technology should conduct the interviews.

[40] Michael Rufo, Itron, notes that "A focus on program induced early replacement versus the effect on efficiency level is gaining attention in the evaluation field. In cases where there is early replacement, two net savings components may be needed to appropriately characterize overall net savings: (1) the early replacement period that uses an in-situ baseline; and, (2) the efficiency increment above minimum or standard practice at the end of the early adoption period (that is, one for the RUL (remaining useful life) period and one for the remainder of the EUL [effective useful life]."

- Establish *a priori* rules for treatment of missing/don't knows in the scoring algorithm.
- Weight the estimates by annual savings to account for the size of the savings impacts for each consumer.
- Sample, calculate, and report the precision[41] of the estimate for the design element of interest (measure, project type, or end use).
- Conduct sensitivity testing of the scoring algorithm.
- Define what the spillover measurement is and is not attempting to estimate and justify the use of an approach.
- Employ, where feasible, a preponderance of evidence (or triangulation of results) approach that uses data from multiple sources (see Itron, Inc. 2010), especially for large savers and complex decision-making cases. Potential data sources could include project file reviews, program staff and account manager interviews, vendor interviews, and observations from site visits.

The New York Department of Public Service (2012) developed additional guidelines specific to the estimation of spillover savings to address recurring methodological limitations that the New York Department of Public Service staff and its contractor team observed in the estimation of spillover in New York and the industry as a whole. Prahl et al. (2013) summarize this work and the critical decisions that evaluators must make before deciding whether and how to estimate spillover. That paper also discusses how the estimation of per-unit gross savings, estimation of program influence, and documentation of causal mechanisms varies for different levels of rigor.

4.3.2 Surveys of Program Nonparticipants

Self-report surveys with nonparticipants are commonly used to triangulate participant self-report responses and collect data for calculating nonparticipant spillover or market effects. These surveys help evaluators understand what EE actions nonparticipants have taken and whether they took those actions because of program influences (nonparticipant spillover). Conducting surveys with nonparticipants poses its own unique challenges:

- There is no record of the equipment purchase, and identifying a group of nonparticipants who have installed energy-efficient equipment on their own can be time consuming and costly.[42]
- Establishing causality entails estimating gross unit savings (often with limited evidence other than the consumer self-report) and establishing how the program may have influenced the consumer's decision. The consumer may not have been aware, for example, of the influence the program had on the equipment's availability or the market actor's stocking practices.

[41] The New York Department of Public Service (2013a) presents guidelines for calculating the relative precision of program net savings estimates for different types of estimates, including the NTG ratio based on the self-report method and for spillover savings. Additional discussion of sampling for evaluation can be found in Khawaja et al. (2013).
[42] One approach to mitigating the efficiency and cost of this is to use one nonparticipant survey that asks about a variety of program eligible measures and use the results across multiple programs.

4.3.3 Market Actor Surveys

When estimating net savings, it is important to consider all the points of program influence. In addition to targeting consumers, upstream and midstream programs often target program services and/or funding to market actors (such as contractors, auditors, and design specialists) with the goal of influencing their design, specification, recommendation, and installation practices. In upstream and midstream programs, consumers may not be aware of program influences on sales, stocking practices, or prices (discussed in the Appendix).[43] Thus, using only participant self-reports when estimating net savings is inappropriate. In these cases, evaluators use market actor self-report surveys to examine the effects of these upstream influences.

These market actor self-report surveys can be designed as qualitative in-depth interviews or as structured surveys with a statistically designed sample of contractors. The use and application of the data determine the format. For example, evaluators may use:

- Qualitative, open-ended data based on a small sample of market actors to contextualize market actors' practices (best used for triangulation purposes).
- Quantitative market actor data to calculate free-ridership and spillover rates specifically related to the practices of those market actors. The calculated rates can then be directly integrated with participant self-report results, triangulated with participant self-report results, and/or used as the sole source for free-ridership and spillover rates. (See, for example, KEMA, Inc. [2010].)

Evaluations can also include market actor survey data to estimate nonparticipant spillover and market effects. An important issue related to the quantification of nonparticipant spillover savings using only surveys of consumers is valuing the savings of measures installed outside the program. As previously noted, during telephone interviews consumers often cannot provide adequate equipment-specific data on new equipment installed either through or outside a program. Although they can usually report what type of equipment was installed, consumers typically cannot provide sufficient information about the quantity, size, efficiency, and/or operation of that equipment to enable a determination about its program eligibility.

One approach to estimating nonparticipant spillover and market effects via market actors is to ask market actors questions such as:

- What percentage of their sales meets or exceeds the program standards for each program measure category installed through the program(s)?

[43] There are studies that focus on examining how a change in the price of an energy-efficient product influences consumer purchases. Two approaches were used: (1) stated preference experiments that systematically ask potential consumers what they would choose from a set of options with different features and prices; and (2) revealed preference studies observe the actual choices consumers make from true choices available to them when making purchases. To obtain accurate revealed preference information, it is usually necessary to observe the items purchased. Consumers cannot reliably report the efficiency levels of recently purchased equipment. Direct observation can be accomplished via store intercepts for small items such as light bulbs, or via onsite visits for large items such as refrigerators. The remaining challenge for this method is the potential nonresponse bias; that is, potential differences between consumers who are willing to have their purchases observed and those who decline. An example of a study that focuses on how changes in price influence consumer purchases of energy efficient products is Cadmus (2012b). See the Appendix for additional information.

- What percentage of these sales did not receive an incentive?

The market actors should then be asked several questions about the program's impact on their decisions to recommend and/or install this efficient equipment outside the program.

4.3.4 Case Studies for Estimating Net Savings Using Survey Approaches

This section presents examples of estimating net savings with self-report surveys. Because self-report surveys are one of the most commonly used approaches, we provide four examples in this section. The first example demonstrates how the participant self-reports method is used to calculate free-ridership of residential and nonresidential programs in Illinois. The second example draws from work in California where self-report surveys are used to estimate free-ridership in nonresidential programs. A third example demonstrates how a sample set of survey questions were used in conjunction with a matrix to estimate free-ridership. The final example summarizes an approach used by the Energy Trust of Oregon (Castor 2012) that calculates low, mid, and high scenario NTG ratios to account for "Don't Know" responses to certain questions. This example addresses the best practice of conducting sensitivity analysis on the algorithm used to estimate NTG.

Example 1. Residential and Nonresidential Programs Free-Ridership Assessment

As part of a literature review for the Massachusetts Program Administrators to examine recent efforts to standardize measurement of net savings,[44] the evaluation team reviewed the recent efforts in Illinois to obtain consistent NTG methods. Below we summarize the background of this effort and the resulting recommended methods in both the residential and nonresidential program areas.

The Illinois (IL) Commerce Commission directed their evaluation teams to compile and formalize consistent NTG methods for use in IL EM&V work. The Commission's directives were twofold: (1) assess NTG methodologies and survey instruments that have been used to evaluate energy efficiency programs, and (2) compile the most justifiable and well-vetted methodologies in an attachment to the updated Illinois Technical Reference Manual (TRM) (Illinois Energy Efficiency Stakeholder Advisory Group 2016). The Commission noted that the IL NTG Methods should be flexible and adaptable to multiple program designs and budgets. It also noted the Methods should be tailored to appropriately assess the specifics of each of the Program Administrators' energy efficiency programs. The resulting statewide NTG methodology document covers the majority of residential and nonresidential programs offered in IL. If the NTG protocol is no longer appropriate, instructions are included for diverging from the IL NTG Methods.

Overview of Residential NTG Approaches—Illinois TRM

The Illinois TRM includes a residential cross-cutting NTG protocol as well as protocols for specific residential programs, including Appliance Recycling, Upstream Lighting, Prescriptive Rebate, Single Family Home Energy Audit, and Residential New Construction. The cross-cutting residential protocol formulates the core NTG as 1 – free-ridership (FR) + participant spillover

[44] Tetra Tech, NMR, and DNV GL, Net-to-Gross Methodology Research, prepared for the Massachusetts Program Administrators, March 24, 2017.

(PSO) and provides specific questions and scoring algorithms for measuring FR and PSO. It also provides specific questions and algorithms for measuring nonparticipant spillover (NPSO) from trade allies and customers, implying that they are to be included in the core NTG formula. This cross-cutting protocol provides detail on measuring PSO and NPSO, but it defers to the specific program protocols for measuring FR.

The specific protocols for programs include:

- The Appliance Recycling protocol includes basic and enhanced self-report approach (SRA) methods with specific questions and scoring algorithms for measuring FR including questions on how the appliance would have been disposed of in the absence of the program. The enhanced method may include additional research methods such as a retailer survey, appliance market assessment survey, or nonparticipant survey. The protocol does not provide specific guidance for when to use each SRA method nor for measuring SO, and thus the cross-cutting protocols may be assumed to prevail.
- The Residential Upstream Lighting protocol recommends using store intercept surveys for the customer SRA to measure PSO and NPSO. The protocol includes specific questions and scoring algorithms for measuring these NTG components. It includes specific questions for measuring FR and allows for partial FR. These include questions to assess program influence (captures the maximum level of program influence, reported by a survey respondent, of the residential lighting program on their decisions to purchase program bulbs on the day of the survey) and no-program questions (used to estimate how many program bulbs a survey respondent would have purchased in the absence of the residential lighting program); FR is calculated as the average of the responses to the two questions.
- The Prescriptive Rebate with No Audit protocol provides basic and enhanced methods with specific questions and scoring algorithms for measuring FR. Questions include program influence and no-program components[45] as well as consistency check questions on the program's influence to resolve possible conflicting responses. The basic method measures FR using a customer SRA. The enhanced method provides a protocol to triangulate and develop a weighted combination of FR estimates from two sources: the basic method and a trade ally survey. When multiple methods are used, evaluators may triangulate results by rating the analysis methodology and data collected using responses (rated on a scale of 0 to 10) to three questions: how likely is the approach to provide a more accurate estimate of FR, how valid is the data collected and the analysis performed, and how representative is the sample. The weight for each method is the average score for that method divided by the sum of the scores for all methods.
- The Single-Family Home Energy Audit protocol provides specific questions and scoring algorithms for measuring FR with different approaches for free/direct install versus rebated/discounted measures. The protocol measures FR using a customer SRA with questions on installation timing, quantity, and no-program scenario. Program influence

[45] Respondents are asked to report their likelihood (using a 0 to 10 scale where 0 is "not at all likely" and 10 is "extremely likely") to implement specified energy efficiency measures in the absence of the program. That likelihood score is then divided by 10 to produce the no-program score.

questions are excluded for free/direct install and included for rebated/discounted measures. It also includes consistency check questions on the program's influence for rebated/discounted measures to resolve possible conflicting responses.

- The Residential New Construction protocol recommends using builder surveys for the participant SRA to measure FR, PSO, and NPSO. The protocol includes specific questions and scoring algorithms for measuring these NTG components. The protocol measures FR using a participant SRA with questions on program influence installation timing, quantity, and no-program scenario. It also includes consistency check questions on the program's influence to resolve possible conflicting responses. PSO includes additional questions to help estimate the amount of savings using IL TRM protocols, such as quantity of appliances or location and amount of insulation. NPSO is based on surveys of two groups: dropout builders not participating in the past 12 months and true nonparticipating builders.

Overview of Nonresidential NTG Approaches—Illinois TRM

The IL TRM includes a core NTG protocol for nonresidential programs as well as protocols for specific programs, including the C&I New Construction, Small Business, and Study-based programs (e.g., programs that include an energy audit or assessment). There are core protocols for FR, PSO, and NPSO that provide specific questions and scoring algorithms associated with calculating FR and SO scores. That said, the core NTG ratio for an energy efficiency program is defined as 1 – FR even though they define PSO and NPSO.

The core FR protocol comprises three scores: Program Components FR Score, Program Influence FR Score, and No-Program FR Score, each ranging from 0 (no FR) to 1 (full FR). The three scores are combined to calculate the FR value. They are calculated as follows:

Program Components FR Score: Participants are asked to rate the importance of various factors on the decision to implement energy efficiency measures. The numeric scales range from 0 to 10, where 0 means "not at all important" and 10 means "extremely important." The factors included in the survey are program and non-program factors that could impact the participant decision-making process. The evaluator can calculate the score in one of two ways:

1. Equal to 1 - ([Maximum Program Factor Rating]/10).
2. Equal to 1 - ([Maximum Program Factor Rating]/([Maximum Program Factor Rating]+[Maximum Non-Program Factor Rating])).

Program Influence FR Score: Respondents are asked to allocate 100 points to the program and to non-program factors. The points the participants allocate to the program are the "Program Points." The "Program Influence FR Score" is calculated as 1 - (Program Points/100).

No-Program FR Score: Respondents are asked to report their likelihood (using a 0 to 10 scale where 0 is "not at all likely" and 10 is "extremely likely") to implement specified energy efficiency measures in the absence of the program. That likelihood score is then divided by 10 to produce this score.

The TRM states that consistency checks should be included in the survey questions to check the consistency of the FR responses. The protocol also provides guidance around vendor influence, including when and how to incorporate vendor responses into the FR calculation. The TRM outlines three scenarios to help decide when to utilize vendor responses, which is based on how involved the trade allies are in the program (i.e., integral in the delivery; part of a select, pre-approved network; implement projects and submit applications on behalf of the customer; sign agreements with the program administrator; or complete program-sponsored training). If vendor surveys are used, the TRM outlines questions that can be asked and based on the responses, when the results would be incorporated. Based on three scenarios, the evaluator decides if the vendor rating should be considered a program factor or non-program factor.[46]

- The Small Business protocol follows the core nonresidential FR protocol but includes a few exceptions primarily to reduce respondent burden.
- The C&I New Construction protocol follows the core nonresidential FR protocol but removes the timing aspect, as the program typically does not impact the acceleration of the construction.
- The Study-based protocol follows the core nonresidential FR protocol but includes additional questions about maintenance and performance of the measure.

The residential cross-cutting protocol states that FR questions should be asked near the beginning of the participant survey, before satisfaction questions. It also states that when estimating SO based on trade ally surveys, respondents should be allowed sufficient time to collect data to inform their responses and not rely on guesses.

The nonresidential core protocol does not provide direction on the timing of the FR survey. However, for SO, the protocol states the PSO module can be implemented as part of the NTG survey or separately, but timed to allow sufficient time—a minimum of three months—after program participation to allow for SO to occur.

Example 2. Nonresidential Programs Free-Ridership Assessment

The Large Nonresidential Free-Ridership Approach, developed by the Nonresidential Net-to-Gross Ratio Working Group for the Energy Division of the California Public Utilities Commission (2012), was developed to address the unique needs of large nonresidential customer projects developed through EE programs offered by the four California investor-owned utilities and other third parties. The Large Nonresidential Free-Ridership Approach is based on an approach that has been evolving for more than 15 years. As described in the framework, the method relies exclusively on the self-report approach to estimate project- and program-level NTG ratios, because the working group notes that other available methods and research designs are generally not feasible for large nonresidential customer programs. This methodology provides a standard framework, including decision rules, for integrating findings from quantitative and qualitative information in the systematic and consistent calculation of the NTG ratio.

The approach describes three levels of free-ridership analysis. The most detailed level of analysis, the Standard – Very Large Project NTG ratio, is applied to the largest and most

[46] Illinois Statewide Technical Reference Manual for Energy Efficiency, Version 5.0. Volume 4: Cross Cutting Measures and Attachments. February 11, 2016, Page 34.

complex projects (representing 10%–20% of the total projects) with the greatest expected levels of gross savings. The Standard NTG ratio, involving a somewhat less detailed level of analysis, is applied to projects with moderately high levels of gross savings. The Basic NTG ratio is applied to all remaining projects.

Five potential sources of free-ridership information are discussed in this study. Each level of analysis relies on information from one or more of these sources:

- **Program files**, which can include various pieces of information relevant to the analysis of free-ridership. Program files may include letters written by the utility's customer representatives that document what the consumer had planned to do in the absence of the rebate and explain the consumer's motivation for implementing the EE measure. It can also include information on the measure payback with and without the rebate.
- **Decision-maker surveys**, conducted with the person involved in the decision-making process that led to the implementation of measures under the program. This survey obtains highly structured responses concerning the probability that the consumer would have implemented the same measure in the absence of the program.
    - Participants are asked about the timing of their program awareness relative to their decision to purchase or implement the EE measure.
    - They are asked to rate the importance of the program versus non-program influences in their decision-making.
    - They are asked to rate the significance of various factors and events that may have led to their decision to implement the EE measure at the time that they did (for example, age or condition of the equipment, information from a facility audit, standard business practices, and experience with the program or measure).

  The survey also asks participants to describe what they would have done in the absence of the program, beginning with whether the implementation was an early replacement action. The decision-makers are asked to describe the equipment they would have installed in the absence of the program, including the efficiency levels and quantities. This information is used to adjust the gross engineering savings estimate for partial free-ridership.

  This survey contains a core set of questions for Basic NTG ratio sites, and several supplemental questions for both Standard and Standard – Very Large NTG ratio sites. For example, if Standard or Standard – Very Large respondents indicate that a financial calculation entered highly into their decision, they are asked additional questions about their *financial criteria* for investments and their rationale for the current project. These questions are intended to provide a deeper understanding of the decision-making process and the likely level of program influence versus these internal policies and procedures. Responses to these questions also serve as a basis for consistency checks to investigate conflicting answers about the relative importance of the program and other elements in influencing the decision. Standard – Very Large respondents may also receive additional detailed probing on various aspects of their installation decision based on industry- or technology-specific issues, as determined by review of other information sources. For

Standard – Very Large sites, the respondent data are used to construct an internally consistent “story” that supports the NTG ratio calculated, based on the overall feedback.

- **Vendor surveys** are completed for all Standard and Standard – Very Large participants who used vendors, as well as for Basic participants who indicate a high level of vendor influence in the decision to implement the EE measure. For participants who indicate the vendor was very influential in decision-making, the vendor survey results are incorporated directly into the NTG ratio scoring.
- **Utility and program staff interviews** for the Standard and Standard – Very Large NTG ratio analyses. Interviews with utility staff and program staff are also conducted to gather information on the historical background of the consumer’s decision to install the efficient equipment, the role of the utility and program staff in this decision, and the names and contact information of vendors involved in the specification and installation of the equipment.
- **Other information** for Standard – Very Large Project NTG ratio sites includes secondary research of other pertinent data sources. For example, this could include a review of standard and best practices through industry associations, industry experts, and information from secondary sources (such as the U.S. Department of Energy’s Industrial Technologies Program’s Best Practices website). In addition, the Standard – Very Large NTG ratio analysis calls for interviews with other employees at the participant’s firm, sometimes in other states, and equipment vendor experts from other states where the rebated equipment is installed (some without rebates) to provide further input on standard practice within each company.

Table 4 shows the data sources used in each of the three levels of free-ridership analysis. Although more than one level of analysis may share the same source, the amount of information used in the analysis may vary. For example, all three levels of analysis obtain core question data from the decision-maker survey.

Table 4. Information Sources for the Three Levels of NTG Ratio Analysis

	Program File	Decision-Maker Survey Core Question	Vendor Surveys	Decision-Maker Survey Supplemental Questions	Utility and Program Staff Interviews	Other Research Findings
Basic NTG ratio	√	√	√[a]		√[b]	
Standard NTG ratio	√	√	√[a]	√	√	
Standard NTG ratio—Very Large Projects	√	√	√[c]	√	√	√

[a] Performed only for sites that indicate a vendor influence score greater than maximum of the other program element scores.
[b] Performed only for sites that have a utility account representative.
[c] Performed only if significant vendor influence is reported or if secondary research indicates the installed measure may be becoming standard practice.

Example 3. Free-Ridership Assessment for an Equipment Rebate Program

This example shows how to calculate an NTG ratio and how to use a sample set of survey questions in conjunction with a matrix to estimate free-ridership (see Table 5). The example is from Chapter 5 of the Energy Efficiency Program Impact Evaluation Guide (SEE Action 2012b). In this case, the evaluators assign a free-ridership score based on a participant's response to six questions.

Table 5. Example Assignment of Free-Ridership Score Based on Participant Responses

Free-Ridership Score	Already Ordered or Installed	Would Have Installed Without Program	Same Efficiency	Would Have Installed All the Measures	Planning to Install Soon	Already in Budget
100%	Yes	Yes	—	—	—	—
0%	No	No	—	—	—	—
0%	No	Yes	No	—	—	—
50%	No	Yes	Yes	Yes	Yes	Yes
25%	No	Yes	Yes	Yes	No	Yes
25%	No	Yes	Yes	Yes	Yes	No
0%	No	Yes	Yes	Yes	No	No
25%	No	Yes	Yes	No	Yes	Yes
12.5%	No	Yes	Yes	No	No	Yes
12.5%	No	Yes	Yes	No	Yes	No
0%	No	Yes	Yes	No	No	No

*Source: SEE Action (2012b) based on example provided by Cadmus.

One issue with this method is the somewhat arbitrary nature of assigning free-ridership scores based on sets of question responses, as they depend on the judgment of the evaluator. Different researchers may assign different free-ridership scores to different sets of respondent answers. To address this, the literature recommends using sensitivity analyses around the free-ridership scores, based on the judgments of people familiar with the program.[47] An example of increasing the robustness of this method is found in an assessment of residential heating and cooling equipment for the Electric and Gas Program Administrators of Massachusetts.[48] Another useful exercise is to assess the reliability of the assignment of free-ridership scores by the evaluators. Inter-rater reliability scores[49] can be calculated to assess the reliability of these assignments. To the extent that evaluators assign the same free-ridership scores to the same set of response

[47] Issues may arise if these free-ridership scores are viewed as categories rather than as continuous variables. A 50% score may imply a higher level of free-ridership than does a 25% score, but it may not denote that the 50% score implies that free-ridership is, in fact, twice as high compared to respondents placed in 25% free-ridership score category. It is possible to perform arithmetic on these numbers and use the values to generate a mean value and even a variance, but this may not be appropriate. The lack of an accurate "distance" factor in these numbers makes the calculated variance hard to interpret. For variables that are meant to represent categories rather than continuous numeric values, frequencies are the more often used descriptive statistic.

[48] This work was conducted by a consortium of consultants under a prime contract led by Cadmus, supported by Navigant, and Opinion Dynamics Corporation (cited as Cadmus; Navigant Consulting; Opinion Dynamics Corporation (2012).

[49] *Inter-rater reliability*, *inter-observer reliability*, and *inter-judge agreement* are some terms that have been used in the literature to designate a wide variety of concepts. All these terms, however, refer to the extent of agreement among raters, judges, and observers (Gwet 2010, 2012).

patterns, then reliability will be increased. Other approaches use upper and lower bounds on free-ridership developed directly from survey respondents.[50]

Example 4. Commercial, Industrial, and Residential Scenario Analysis

The Energy Trust of Oregon uses an approach (Castor 2012) to calculate low, mid, and high scenario NTG ratios to account for the "Don't Know" responses to certain questions. The report appendix describes this approach. The project's free-ridership score is composed of two elements: a project change score and an influence score.

The project change score is based on the respondent's answer to the question, "Which of the following statements describe the actions you would have taken if Energy Trust incentives and information were not available"? Possible answer choices are assigned a number between 0 and 0.5, with 0 indicating no free-ridership and 0.5 indicating that the participant was a full free-rider. Because a respondent can select multiple responses to the question, the answer choice with the lowest score is selected. If the respondent selects "Don't Know," two scores are created to account for the range of possible answers (0 and 0.5).

For commercial projects, respondents are asked this follow-up question when they report they would not have done anything differently in the absence of the program: "If your firm had not received the incentive, would it have made available the funds needed to cover the entire cost of the project"? If the respondents select "Yes," their project change score is 0.5. If the respondents select "No," their project change score is 0. However, if the respondents select "Don't Know," they are given two scores for project change, as previously described.

The influence score is based on respondents' answers to questions about the influence of Energy Trust incentives, program representatives, contractor/salesperson, studies, and other program elements. The answer choices are given a value between 0 (element's influence was a 5, extremely influential) and 0.5 (element's influence was a 1, not at all influential). The score for the most influential element is taken as the influence score. If respondents answer "Don't Know" for all elements, they are given two influence scores to account for the range of possible answers (0 and 0.5).

---

[50] Violette et al. (2005) discuss approaches used in the net savings and attribution assessment for a large-scale C&I retrofit program. Free-ridership was assessed using a series of survey questions asked of various actors, including participating end-use consumers and vendors/contractors/consultants. Free-ridership was asked in direct free-ridership questions and supporting, or influencing, questions. Participating owners and energy service companies/contractors in a large-scale C&I retrofit program were each asked for direct estimates of: (1) the "proportion" of the savings or measures that would have been installed without the program; and (2) the "likelihood" that the measures would have been installed without the program. A three-step approach was used. Step 1 focused on whether the respondent believed that free-ridership existed at all; if the respondent believed it existed in this project, Step 2 established bounds on the free-ridership effect, that is, what was the smallest value that seemed reasonable and what might have been the highest reasonable free-ridership value. Step 3 used questions to obtain where within this range the free-ridership value was likely to fall. Appendices to Violette et al. (2005) discuss alternative approaches. This program had some unique characteristics that made this approach more tractable. It involved large-scale C&I projects and the survey respondents were provided with summaries of the technologies and measures installed. Other efforts that used similar approaches include Violette, Ozog and Cooney (2003) for addressing net savings from regional and market transformation programs in the Pacific Northwest, and Navigant (2013b) which assesses the net impacts of U.S. DOE's Wind Powering America Initiative.

To generate the free-ridership score for each project, the project change and influence scores are added. For respondents who do not provide "Don't Know" answers, this score will be a single number between 0 (no free-ridership) and 1 (full free-ridership). For those who gave a "Don't Know" answer to one of the questions, there are two free-ridership scores—one high and one low. For those who answered "Don't Know" to *both* the project change and influence questions, no score is calculated.

Free-ridership scores are averaged for all respondents in each program/measure group and the result is shown as a percentage rather than a decimal (see Table 6 for pros and cons of survey-based approaches).

- "Low Scenario" is the average of the free-ridership scores where the low score is used for those who answered "Don't Know" to a question.
- "High Scenario" is the average where the high score is used for those who answered "Don't know" to a question.
- "Mid Scenario" is the average of the Low and High Scenarios. In the case of C&I projects, individual scores are weighted by their share in the electricity or gas savings of all respondents of their group before the scores are averaged for scenarios.

Table 6. Survey-Based Approaches—Summary View of Pros and Cons

Pros	Can provide useful information to support process and impact evaluations (for example, source of awareness, satisfaction, and demographics) Flexible approach that allows the evaluator to tailor questions to the program design or implementation methods Participant self-reports can yield estimates of free-ridership and spillover without the need for a nonparticipant control group Nonparticipant and market actor interviews can be used to triangulate participant self-report responses and calculate nonparticipant spillover or market effects.
Cons	Potential biases related to respondents' giving "socially desirable" answers Consumers' inability to know what they would have done in a hypothetical alternative situation, especially in current program designs that use multiple methods to influence behavior The tendency of participants to rationalize past choices Potential arbitrariness of scoring methods based on evaluator judgment that translate responses into free-rider estimates Participants may fail to recognize the influence the program may have had on other parties who influenced their decisions (for example, program may have influenced contractor practices, which in turn impacted the participant) Participant surveys capture only a subset of market effects Amount of time and cost to identify a group of nonparticipants who have installed energy-efficient equipment on their own.

4.4 Market Sales Data Analyses (Cross-Sectional Studies)

A market sales data method can capture the total net effect of the program, including both free-ridership and participant and nonparticipant "like" spillover. As described in a residential free-ridership and spillover methodology study prepared for the Massachusetts Program Administrators (NMR Group, Inc. and Tetra Tech 2011), the total net effects of a program can be estimated via an analysis of market sales data.

The most common approach is a cross-sectional comparison area method in which post-program data are compared with data from a non-program comparison area (or multiple comparison areas) for the same point in time. Thus, evaluators can make a comparison between the change in the program area from the pre-program period to the post-program period *and* the change in the non-program area over the same period.

The NMR Group, Inc. and Tetra Tech (2011) study lists three important factors to consider when deciding if an approach is appropriate for a program:

- **Does an appropriate comparison area exist?** Comparison area(s) must represent a credible baseline for the area of interest. This may entail using a set of systematic adjustments to control for differences in total size of, or demographics for, the areas. As EE programs become more prevalent, finding comparison areas that do not have similar program activities is becoming more difficult.
- **Are the market data available and complete?** Market data analysis requires comprehensive market data for the area of interest and an appropriate comparison area or areas. The complication here is that comprehensive sales/shipment tracking systems have not been available for most markets. Absent comprehensive sales data, a general picture of market coverage can be obtained by conducting surveys or in-depth interviews. These are typically conducted with vendors and contractors about sales volumes and efficient equipment sales shares for conditions with and without the program, or for in-territory and comparison area sales. In some cases, the self-reported purchases of participating end users can provide market data if the sample is sufficiently large and representative of the market. Also, it can be expensive to gather the market sales and shipment data, and even a diligent data collection effort may leave gaps in the data.
- **What are the features of the program?** Market data analysis is usually appropriate for programs that promote large numbers of homogenous measures and that have substantial influence upstream to the end user.

As an example of this approach, Cadmus et al. (2012) tracked ENERGY STAR® appliances, lighting, and home electronics product sales in New York and then compared those sales to sales of the same products in Washington, D.C., Houston, Texas, and Ohio. All these baseline areas were without significant utility efforts to promote ENERGY STAR products. The market data were used to estimate both the market share and the energy savings attributable to the New York Energy $mart Products Initiative Program administered by the New York State Energy Research and Development Authority.[51]

[51] Scott Dimetrosky indicated that this study developed savings from product sales and installations. These savings were derived by first estimating the market share for ENERGY STAR products through estimates of total market size and sales of ENERGY STAR products. Next, portions of the market share were allocated to exogenous, non-New York Energy $mart Products Initiative Program (NYE$P) effects, including the impact of the national U.S.

Another example of a market sales approach entails interviewing or surveying a panel of trade allies who are either program participants or nonparticipants. This could include contractors, retailers, builders, and installers. These trade allies are offered monetary compensation for information about projects or sales completed within a specified time period (see Table 8for pros and cons of this approach). The types of information requested can include manufacturer, efficiency levels, size, price, installation date, installation ZIP code, types of incentives received, and an assessment of the program's impact on incented and non-incented efficiency actions. With annual updates, this method could provide context for tracking longer term ongoing program impacts or market effects. This method could also work in tandem with other approaches for estimating net savings and provide a market context for estimates that may otherwise focus only on short-term impacts.

Another more detailed example of a recent market sales data analysis using in-store visits and web scraping is shown below.

4.4.1 Case Study for Market Sales Data Analysis

Example 1: Massachusetts RLPNC 16-6: Lighting Shelf Stocking

On behalf of the Massachusetts ENERGY STAR Lighting program administrators and Lockheed Martin, NMR Group, Inc. conducted a shelf-stocking and price survey to evaluate the impact of the Mass Save® residential lighting program on consumer retail lighting in Massachusetts (NMR 2017).

The study took advantage of two separate but complementary data collection methods: 1) site visits to 100 stores in Massachusetts and 30 stores in New York in 2016 to inventory light emitting diode, compact fluorescent lamp (CFL), incandescent, and halogen lamp packages, and 2) web scraping, which provided time series data on lighting cost and availability through the collection of data from retailer web pages. The authors noted that while shelf-stocking studies provide a useful look at lighting cost and availability at a discrete point in time, web scraping adds time series data on lighting cost and availability in the marketplace over time. Because the study used two methods, the authors could compare the data collected through both methods and learn how online and in-store prices and availability differed.

The authors noted that both methods offered distinct advantages. Physically visiting stores is the only way to learn information about how products are displayed in the store, the amount of shelf space given to different products types, and what indirect and direct signals stores are providing to customers about the value and desirability of the products. Web scraping offered a number of other advantages:

- Eliminated the financial and time cost of travel, training, and obtaining permission to visit retailers

Environmental Protection Agency/U.S. Department of Energy ENERGY STAR Program, naturally occurring adoption (including the impact of higher energy prices and interest generated by programs in neighboring states), and the impacts of other New York State Energy Research and Development Authority residential programs. The remaining market share, after netting out these other effects, was considered attributable to the NYE$P.

- Not particularly difficult or expensive to set up as the standard methods use free open-source software that is widely used and well documented
- Scraping can be easily automated to run on a regular schedule to create rich time-series datasets.

On the other hand, web scraping had some inherent caveats and limitations:

- The information available is only as good as the websites' administrators make it
- Websites tend to change frequently, which requires updates to the code
- Markdown and rebate information is included inconsistently
- There is not a way to verify how online products offerings, prices, and stock data correspond to what is actually in a store
- The amount of data generated results in the need to filter and clean the data to generate useful insights.

The overall conclusion of the study was that incorporating both data streams offered a richer picture of the lighting market during the study period.

The pros and cons of market sales data analyses are listed in Table 7.

Table 7. Market Sales Data Analyses—Summary View of Pros and Cons

Pros	Can estimate the total net effect of a program Uses information on actual consumer behavior Addresses trends in an entire market Most appropriate for programs that promote large numbers of homogeneous measures and have substantial influence upstream.
Cons	There may be a low availability and quality of sales and shipment data in the area of interest and in an appropriate comparison area(s) Data may be expensive to acquire and/or may have gaps that can be misleading May be difficult to determine the appropriateness of a comparison area.

4.5 Structured Expert Judgment Approaches

Structured expert judgment approaches involve assembling a panel of experts who have a good working knowledge of the technology, infrastructure systems, markets, and political environments. This approach is one alternative for addressing market effects in different end-use markets. These experts are asked to estimate baseline market share for a measure or behavior. In some cases, they are also asked to forecast market share with and without the program in place. Structured expert judgment processes use a variety of specific techniques to ensure that the panel of experts specify and take into account key known facts about the program, the technologies supported, and the development of other influences over time (Tetra Tech et al. 2011).

The Delphi process is the most widely known technique (NMR Group, Inc. and Research Into Action 2010). Each panelist is asked to make a judgment on the topic—based on the provided information and on his or her experience—and submit the information to the evaluators. The evaluators compile the information from the panelists and return it to the panelists for another review. The panelists are asked whether they stand by their original judgments or whether the assessments of their peers have caused them to alter their judgments. At least two rounds of judgment are required for a Delphi panel, although more rounds can be used.

Some advantages of the structured expert judgment approach are:

- The estimate is based on feedback from a group of experts, which can be particularly useful for programs with complex end uses.
- It is a useful tool for consolidating results from multiple methods to develop a consensus estimate (see example 2 below).

As with other approaches (such as market sales data analysis), the structured expert judgment method relies on high-quality data to inform the panel, so sparse data can result in inaccurate estimates of net savings (NMR Group, Inc. and Research Into Action 2010).

Two examples of using the structured expert judgment approach to estimate net savings are presented here. The first example describes how Delphi panels were used to estimate net savings for a residential new construction program in California. The second example describes the development a final estimate using a Delphi panel's review of estimates.[52]

4.5.1 Case Studies for the Structured Expert Judgment Approach

Example 1: Residential New Construction Delphi Panel

In a study prepared for the California Public Utilities Commission Energy Division, evaluators used two Delphi panels of Title 24 consultants and building industry experts to convert the gross savings estimates. The panel converted estimates from investor-owned utilities (IOU) programs targeting the residential new construction sector to net savings estimates (Hoefgen et al. 2011).

The panelists received detailed data pertaining to code compliance, compliance margins, and estimates of annual gross energy savings in non-program homes at the state level and by climate region. After reviewing these data, panelists were asked to:

- Estimate the proportion of the electricity and natural gas savings attributable to the IOU programs targeting the residential new construction sector and other factors (non-IOU residential new construction programs, the economy/housing market, energy prices, and climate change).
- Estimate the percentage of net savings in non-program homes attributable to different IOU program elements (builder trainings, incentives, and design assistance).

[52] An application of the Delphi technique as applied outside of EE may be informative. Navigant (2013b) conducted an evaluation of the Wind Power America program. The goal was to assess the impacts attributable to the program. The unique aspect of this Delphi exercise was the use of range estimates; that is, experts were asked about lower and upper bounds to the effects as well as a best estimate. This approach allowed the experts to provide their own insights into the uncertainty of the estimates. Gauging uncertainty and then using that in probabilistic and scenario analyses are consistent with other utility resource planning activities. Adapting these methods to EE resource assessment may increase the usefulness of the information.

- Assess the extent to which the market effects were likely to persist in the absence or reduction of the IOU programs.
- Estimate the percentage of homes that would have been below code in the absence of the IOUs' programs and other factors, and estimate the compliance margin of the below-code homes in the absence of each factor.

Each panelist completed two rounds of detailed surveys. In the second round, they were provided a comparison with other panelists' responses and logic and allowed to change their answers. The evaluation team analyzed the Title 24 consultant responses (both weighted and unweighted) using the building industry experts' responses as a qualitative check. The Delphi panel provided estimates on gross electricity and gross natural gas savings from above-code homes. Both panels identified the various elements of training (builders, subcontractors, and Title 24 and code officials) as the most important elements of the IOUs' programs.

Example 2: Lighting Program Delphi Panel

Another way to use a Delphi panel is to have the panel review estimates derived through other methods to develop a final estimate. As part of the evaluation of the Massachusetts ENERGY STAR Lighting Program (KEMA 2010), evaluators used a Delphi panel of lighting and EE experts across the United States and Canada. The panelists were asked to integrate results from five methodologies that yielded NTG estimates (conjoint analysis, multistate modeling, revealed preference study, supplier interviews, and a willingness-to-pay study). Evaluators then used the Delphi panel's review in developing recommendations for the final NTG estimate.

See Table 8 for pros and cons of the structured expert judgment approach.

Table 8. Structured Expert Judgment Approaches—Summary View of Pros and Cons

Pros	The resulting estimate is the independent, professional judgment of a group of technology and/or market experts It is a useful approach for programs with diverse and complex end uses or practices Is a useful tool for consolidating results from multiple methods to develop a consensus estimate Panel members can provide levels of confidence and procedures using appropriate elicitation methods.
Cons	The approach relies on high-quality data to inform the panel, leading to reasonable estimates of net savings Sampling-based calculations of confidence and precision are not available The approach is judgmental/subjective.

4.6 Deemed or Stipulated Net-to-Gross Ratios

Deemed or stipulated NTG ratios are predetermined values and do not rely on a calculation-based approach. Deemed values are often based on previous NTG research that was conducted using at least one of the other methods described in this chapter.

NTG ratios are often stipulated when the expense of conducting NTG ratio analyses cannot be justified or when the uncertainty of the potential results is too great to warrant a study. A recent review of 42 jurisdictions in the United States and Canada (which represented nearly all jurisdictions with ratepayer-funded EE programs) found that only 14% use a deemed approach to NTG for C&I programs compared to 50% of the jurisdictions using an active research approach to developing estimates of net savings factors (Navigant 2013a).[53]

Deemed or stipulated NTG ratios are typically either set by a regulatory agency or negotiated between regulators and program administrators. These ratios may be determined at the portfolio level (for example, Michigan and Arkansas)[54] or on a measure-by-measure basis (for example, California and Vermont).[55] Typically, evaluators base the ratios on NTG studies from past evaluations and/or reviews of other similar programs in which a NTG ratio was estimated. For example, it is not unusual in a multiyear portfolio cycle to estimate a NTG ratio for an initial year (or possibly every other year), with deemed values used in the subsequent or intervening years. This multiyear estimation of NTG ratios is a compromise between performing net savings estimation studies every year and the use of deemed values based on that research for a selected time period. As an example, Massachusetts moved to this approach.[56]

In other cases, evaluators use historical data or other information from a wide range of sources to develop a "weight of evidence" conclusion about the program's influence (SEE Action 2012b). As discussed earlier, one common approach for developing a stipulated value is to use a panel of experts who have the relevant experience to make that judgment (Delphi panel).

Although using deemed or stipulated values is a relatively simple and low-cost approach, there are several disadvantages. NTG values are variable across time and space, and strongly linked to program design and implementation. This makes deemed values or assumptions potentially unreliable when transferred from a program in one jurisdiction to a similar program in another jurisdiction.[57] NTG values based on primary research efforts can produce estimates that are based on program-specific information (NMR Group, Inc. and Research Into Action 2010). As a

---

[53] Approximately one third of the jurisdictions did not adjust gross savings for either free-ridership or spillover; however, many of those states conducted some NTG research to inform future program design. This reflects policy decisions in each state. Several states that did not adjust gross savings for net savings factors at the time of this study have changed or are contemplating changing to approaches that do estimate net savings. Pennsylvania and Maryland fall into this category. In Pennsylvania, Act 129 program savings targets are based on gross savings, but utilities participating in Act 129 programs are required to report gross and net savings.

[54] Arkansas: NTG deemed at 0.8, www.apscservices.info/pdf/07/07-085-tf_286_44.pdf; Michigan: NTG is deemed at 0.9 for all programs except pilot, education, and low-income programs, which are deemed at 1.0. http://efile.mpsc.state.mi.us/efile/docs/17138/0009.pdf. Note that most low-income programs are not subject to NTG analysis (that is, are deemed at 1.0).

[55] California Database for Energy Efficient Resources (DEER): http://www.deeresources.com/; Vermont, see: www.efficiencyvermont.com/docs/about_efficiency_vermont/annual_reports/2011_Gross_to_Net_Report_EfficiencyVermont.pdf.

[56] Massachusetts has been conducting extensive NTG research, but has moved to deemed/stipulated values for their 3-year plan. Any NTG variances from the stipulated values have no effect on current cost recovery or incentive payments. Yet the extensive program- and measure-level NTG research continues where appropriate, and the state is benefiting from improved program designs without major controversy involving cost recovery and incentives for current programs.

[57] Another issue raised by a reviewer was that the use of deemed NTG values can remove the incentive for the program administrator to reduce free-ridership and maximize spillover and market effects to yield greater net savings values.

result, these values provide useful information for the design and implementation of programs[58] and may mitigate the risk to ratepayers from utilities receiving performance incentive payments on savings not actually attributable to the program (as well as the risk to ratepayers of making performance incentive payments that are too large). NTG values are also critical from a resource planning perspective and having better data on the actual energy savings achieved from energy efficiency programs can help the planning process (Navigant 2013a). Deemed or stipulated NTG values do not provide these benefits.

The following example illustrates how one agency uses deemed savings for program planning.

4.6.1 Case Study for Using Deemed Savings

Example 1: California Public Utilities Commission Database for Energy Efficient Resources

The California Public Utilities Commission uses deemed savings (listed in its Database for Energy Efficient Resources) for planning purposes and interim savings estimates for its programs. These deemed savings are updated based on results of NTG studies. NTG savings values are presented for kilowatt-hours and kilowatts.

See Table 9 for pros and cons of a deemed savings approach.

Table 9. Deemed or Stipulated Approaches—Summary View of Pros and Cons

Pros	This approach can reduce contentious after-implementation adjustments to estimated program savings because agreed-on net savings factors are developed in advance of program implementation.
Cons	An incorrect estimate can be deemed It is not based on current program-specific information The evaluator cannot assign sample-based statistical precision to the estimate Developing deemed savings net values at the measure and technology levels can be time consuming and expensive The process for developing deemed net savings can be contentious.

4.7 Historical Tracing (or Case Study) Method

This method involves reconstructing the events (such as the launch of a product or the passage of legislation) that led to the outcome of interest. An example of this is developing a "weight of evidence" conclusion about the specific influence a program had on the outcome.

Historical tracing relies on logical devices typically found in historical studies, journalism, and legal arguments (Rosenberg and Hoefgen 2009). These include:

- Compiling, comparing, and weighing the merits of narratives of the same set of events provided by individuals who have different points of view and interests in the outcome

[58] For example, free-ridership can inform decisions to discontinue incenting certain measures, increase incentive amounts, or increase the efficiency level being incented.

- Compiling detailed chronological narratives of the events in question to validate hypotheses regarding patterns of influence
- Positing a number of alternative causal hypotheses and examining their consistency with the narrative fact pattern
- Assessing the consistency of the observed fact pattern with linkages predicted by the program logic model
- Using information from a wide range of sources (including public and private documents, personal interviews, and surveys) to inform historical tracing analyses.

The historical tracing method traces chronologically a series of interrelated events either going forward from the research point of interest to downstream outcomes, or working backward from an outcome along a path that is expected to lead to precursor events. If all likely paths are followed, forward tracing can capture a relatively comprehensive view of project or program effects. Because the path leads from a program event, the connection to the event is assured. Backward tracing usually focuses on a single outcome of importance and follows the trail back through developments that seem to have been critical to reaching the identified outcome. These developments may or may not link back to the research program of interest (see Ruegg and Jordan 2007).

Weiss (1997) suggests historical tracing is similar to theory-driven evaluation and can be viewed as an alternative to classical experimental design. This approach suggests that if the predicted steps between an activity and an outcome can be confirmed in implementation, this matching of the theory to the observed outcomes will lend a strong argument for causality. In other words, if the evaluation can show a series of microsteps that lead from inputs to outcomes, causal attribution, for all practical purposes, is supported by this approach.

Scriven (2009) argues that some researchers have been entranced by the paragon of experimental design—the RCT—and have generalized this into a virtual standard for good causal investigation. This view can be contrasted to the way that "epidemiology, engineering, geology, field biology, and many other sciences establish causal conclusions to the highest standards of scientific (and legal) credibility" (p. 151).

This method is best suited to an attribution analysis of major events, such as adoption of new building codes or policies. It is not typically applicable to EE programs. However, various elements of this approach may be used in the analysis of very large custom projects that essentially require case study approaches.

Because this method draws from multiple information sources, it is difficult or impossible to determine the magnitude of the effects, so the evaluator cannot assign statistical precision to the estimate (NMR Group, Inc. and Research Into Action 2010). However, as part of making a persuasive case for attribution and providing evidence supporting a statistically derived net savings estimate, this method can be very important. Statistics alone often do not constitute a complete attribution assessment. They often require context using supporting logic to enhance the validity of the statistical estimates, as illustrated in the following example.

4.7.1 Case Study for Using the Historical Tracing Method

Example 1. Historical Tracing for a Residential New Construction Program

Keneipp et al. (2011) used historical tracing in conjunction with Delphi panels to develop energy savings for new homes (see Table 10 for pros and cons of this approach). This study used historical tracing spanning 14 years of regulatory documents to create timelines of the residential new construction program presence and activities for Arizona Public Service Company. The evaluators used these data to create an influence diagram of market influences on specific building practices. This information was then shared with two in-person Delphi panels of market experts who estimated the percentage of homes built in 2010 using specific building practices. These Delphi panels also developed the counterfactual scenarios used to show the net impact of the residential program on the percentage of homes that were built to standards, but would not have met these standards in the absence of the program. The Delphi outputs were then used to develop inputs for an engineering simulation model to calculate energy savings per home. This example illustrates how historical tracing can be used in combination with other methods to develop actual quantitative net savings estimates from an EE program.

Table 10. Historical Tracing (or Case Study) Method—Summary View of Pros and Cons

Pros	Draws from multiple information sources Can be used at a market level for upstream EE programs Can be useful for making a persuasive case for attribution and provide evidence to support a statistically derived net savings estimate.
Cons	It can be difficult to translate the influence factors into estimates of impacts without additional modeling The evaluator cannot calculate sample-based statistical confidence and precision levels for the estimate.

4.8 Common Practice Baseline Approaches

The common practice baseline approach[59] is also is receiving attention as a method for estimating net savings. SEE Action (2012b) has defined the common practice baseline as follows:

> *Common practice baselines are estimates of what a typical consumer would have done at the time of the project implementation. Essentially, what is "commonly done" becomes the basis for baseline energy consumption (SEE Action, 2012b, p. 7-2).*[60,61]

[59] The Common Practice Baseline section gave rise to several comments. Some reviewers did not see this method as parallel to the other methods presented in this chapter, as it focuses on *ex ante* values of the mean of market behavior and does not look at *ex post* information on actions or program participants. In this context, this approach was viewed as more of an *ex ante* deemed net savings approach (see Section 4.7 on deemed NTG values). After considering these comments, the Common Practice Baseline approach was viewed as warranting a separate section due, in part, to the recent attention given this approach to net savings.

[60] SEE Action (2012b) illustrates this "commonly done" baseline using an appliance example. "For example, if the program involves incenting consumers to buy high-efficiency refrigerators that use 20% less energy than the minimum requirements for ENERGY STAR® refrigerators, the common practice baseline would be refrigerators

This baseline includes a "consideration of what typically would have been done in the absence of the efficiency action" (SEE Action 2012b). This approach is under development in several jurisdictions and will certainly evolve in its application. In general, it is based on using available information to develop an *ex ante* estimate of net savings, with limited adjustments based on *ex post* data and analysis. This approach has many appealing qualities, but the tradeoffs need to be clarified, both in terms of potential biases and the real costs associated with this approach.

The common practice baseline method is relatively new in the broader evaluation literature and its application has been somewhat limited; however, the Northwest Power and Conservation Council (NW Council) in the Pacific Northwest has applied a variant of this method for a number of years in estimating *ex ante* net savings.[62] The NW Council continues to evolve this approach with new protocols developed by the Regional Technical Forum (RTF 2012).[63] Ridge et al. (2013) indicate that, in addition to the NW Council, three other jurisdictions are working with variants of the common practice baseline approach: Indiana, Delaware, and Wisconsin (Focus on Energy). In general, these jurisdictions have evaluation guidelines or regulatory framework that allows for the use of common practice baseline variants under certain circumstances, but they also allow for and use survey-based approaches and RCT or quasi-experimental design approaches to estimated net savings for many programs.

4.8.1 Common Practice Baselines—Discussion

As with other net savings approaches, the common practice baseline approach is designed to assess the savings attributable to EE program activities. One advantage claimed for the common baseline approach is that it avoids double counting of free-riders. The concern is that the two-step approach—where (1) gross savings is estimated *ex post* using a baseline that may be similar to "common practice"; and (2) an NTG ratio is applied to the *ex post* gross savings—can double count at least some free-riders (Ridge et al. 2013; Hall et al. 2013). The argument is that the *ex*

---

that consumers typically buy. This might be non-ENERGY STAR refrigerators, or ENERGY STAR refrigerators, or, on average, something in between."

[61] SEE Action (2012b) defines common practice baselines in its glossary as "The predominant technology(ies) implemented or practice(s) undertaken in a particular region or sector." (p. A-4).

[62] Tom Eckman of NW Council indicated that this general approach has been applied in setting deemed savings since the 1980s, and it was designed to fit with the NW Council integrated planning process; that is, it is meant to provide an estimate of the increment of savings beyond what system planners assume for naturally (or currently) occurring efficiency in their demand models. Additional information can be found at the RTF website of the NW Council and in RTF (2012) as well as in the roadmap for the assessment of EE measures (RTF, 2015).

[63] Some reviewers indicated that this double counting problem may be the result of inconsistent program rules as set out by the program administrators and regulators, not an estimation issue. Further, a number of reviewers indicated that rather than over-estimating free-riders, this approach underestimates free-riders due to selection bias (discussed in the main body text below). The RTF guidelines (dated August 15, 2012) sets out the current practice baseline approach most directly in its definition of savings: "Savings is defined as the difference in energy use between the baseline (see section 3.2) and post (after measure delivery) periods, which is caused by the delivery of a measure. The terms "net" or "gross" are intentionally not used to modify the term "savings," as they may conflict with the definition of "baseline," provided in section 3.2. The current practice baseline defines directly the conditions that would prevail in the absence of the program (the counterfactual), as dictated by codes and standards or the current practices of the market. The most important conflict would arise if savings were estimated against a current practice baseline and then those savings were further adjusted by a net-to-gross ratio, where the net-to-gross ratio was the probability that the measure would have been delivered in the absence of program influence." Note that the RTF uses the term *current baseline* rather than *common practice baseline* used elsewhere.

ante estimate of gross savings may be close to net savings without any adjustment for NTG factors such as free-ridership, spillover, and market effects. This view assumes that some of these NTG factors are already accounted for by the process used to produce the *ex ante* gross savings estimates. This emphasizes the need to: (1) understand the derivation of gross estimates as part of the EE evaluation process, and (2) to explicitly set out the assumed counterfactual scenario in both the gross savings and net savings methods used.[64] Taking these two steps avoids the double counting that results in higher-than-appropriate free-ridership estimates.[65]

Massachusetts has recently adopted a gross baseline framework (DNV GL and ERS 2017) designed to be consistent with the ISO-NE Forward Capacity Market requirement, where baseline is based on the more stringent of existing code/standard and "Industry Standard Practice." The transition plan to implement the new framework includes revising the net-to-gross survey process to ensure that net savings is neither over- nor under-estimated as a result of the gross baseline revision.

Examples from guidelines on common practice baselines include:

- **NW Council's guidelines savings estimation methods:** The NW Council through its Regional Technical Forum (RTF) has the longest history in using a common practice baseline approach. Termed "current practice baseline" by the RTF, this baseline defines directly the conditions that would prevail in the absence of the program (the counterfactual scenario), as dictated by codes and standards or the "current practices of the market." (RTF 2015, p. 3) with current practice defined as the "typical choices of eligible end users, as dictated by codes and standards and the current practices of the market." The RTF estimates this baseline based on recent choices of eligible end users in purchasing new equipment and services. These choices may be inferred from data on shipments, purchases (equipment or services) or selected design /construction features. For example, the baseline for more efficient televisions is the average efficiency of recent television shipments. These baselines along with the measure unit energy savings are subject to a sunset date. The sunset date is "shortened as needed to reliably estimate savings for a measure whose baseline is rapidly changing." (p.10). The RTF sets out indicators used to determine if current practice is the appropriate baseline. However, "as a general rule, the RTF will use a baseline that is characterized by current market practice or the minimum requirements of applicable codes or standards, whichever is more efficient." (p.10).

[64] It is important to remember that both gross savings and net savings are difference estimates and both need a baseline for estimation (see NEEP, 2016).
[65] Some reviewers indicated that this double counting problem may be the result of inconsistent program rules as set out by the program administrators and regulators, and is not an estimation issue. If this is the case, evaluators still must decide whether the *ex ante* savings are net, gross, or somewhere between, because the *ex post* estimates must be used in an internally consistent way to adjust the claimed *ex ante* savings. Further, a number of reviewers indicated that rather than overestimating free-riders, this approach is likely to underestimates free-riders because of selection bias (discussed in this section).

Indiana and Delaware evaluation frameworks: The evaluation guidelines developed in two state-wide frameworks[66] list the use of the standard market practice as approaches that can be used in the estimation of net savings in utility evaluation of EE programs. Indiana indicates that this approach is a way to set energy impact analysis baselines so that the baseline already incorporates the influence of free-riders. In this approach, a free-rider assessment is not needed because the market is already using a standard market practice baseline without the program's direct influence. This baseline is typically set at the mean of the level of EE being installed across the market being targeted by the program (TecMarket Works et al. 2012, p. 55). An update to the Delaware State-Wide Evaluation Framework (Optimal Energy, 2015) also listed standard market practice baselines as a candidate approaches for use in estimating net savings. "Because free-riders are expected to take part in Delaware programs, a Net-to-Gross analysis will be completed for all programs in which free-riders are expected, unless the evaluation approaches use experimental or quasi-experimental designs or set energy impact baseline conditions at standard market practice levels that lead directly to the estimation of net savings." In addition to the evaluation guidelines discussed above, the Wisconsin Focus on Energy (FOE) used a common practice baseline method for a residential program in recent evaluation work (FOE 2017) and sets out processes for use of this method in future evaluation work. The case for the use of a common practice baseline approach appears to stem largely from two issues:

1. The definition of gross savings may include factors that are more appropriately viewed as components of net savings, and additional adjustments are not needed to these original estimates. This is essentially an *ex ante* estimate of net savings using current practice as the baseline with net savings estimated as the reduction in energy use resulting from the change to more efficient technologies.[67,68]
2. Program evaluations that report net savings may do so inconsistently. Unfortunately, the components of the net savings calculation differ between jurisdictions, and are often based on what the jurisdiction's stakeholders view as appropriate and measurable (see

[66] These two state-wide frameworks provide guidance on evaluation methods for utility EE evaluations and include the use of common or standard practice baselines as candidate methods; however, it is not clear how often a common or standard practice baseline method has been selected for use by utilities in these states. An evaluation report addressing Indianapolis Power and Light's EE programs (2015) did not use CPB methods and instead used survey methods for estimating net savings in C&I and residential programs.

[67] Tom Eckman of the NW Council expands on this point, stating that, "What is occurring prior to program launch is a better measure of what would have occurred absent the program (that is, the counterfactual scenario) than a determination made after the program has influenced the market." Essentially, the NW Council performed an *ex ante* net analysis when they developed deemed savings estimates that are by design viewed as net savings. For the NW Council's purposes, this is viewed as being as accurate as performing complex studies after the program has been implemented. More information on the NW Council approach can be found in RTF (2012) and at the RTF website http://rtf.nwcouncil.org/.

[68] The common practice approach as applied by the NW Council works best when the forecasts are made at the measure level. Covering all the measures that combine to make a program can be time consuming and expensive to update. Also, this is short term in that over time, the control group (that is, nonparticipants) would likely have evolved their actions from one year to the next as conditions change and accounting for these effects is important in determining net savings. As with all approaches discussed in this section, there are pros and cons and the selection of the approach to use and the context in which this choice is made influences these decisions. For example, Tom Eckman of the NW Council indicated that this method may be less controversial in the Northwest because some entities do not have financial incentives tied to estimates of net savings.

NEEP 2012). Although spillover is widely recognized and can be significant, several jurisdictions resist estimating spillover values and including them in the net savings calculations. Market effects values have faced similar challenges.[69]

The NW Council and the RTF have used common practice baselines for energy savings more consistently and longer than any other region or jurisdiction. Much of the RTF work is regional which can help define appropriate markets for both residential and non-residential appliances and equipment. In addition, regional organizations in the northwest (e.g., the Bonneville Power Authority[70] and Northwest Energy Efficiency Alliance) conduct market characterizations for important energy using equipment providing information that the RTF can use to develop these baselines. Finally, the RTF supports the NW Council in the development of a regional power plan every 5 years. The use of the energy efficiency baselines by the RTF are designed to be consistent with the assumptions used in the most recent Power Plan. The RTF has considered the context and needs to be met by its savings estimates and has designed these savings approaches to meet these needs.[71,72] It should also be noted that some entities use other methods discussed in this section to estimate net savings. For example, the Energy Trust of Oregon (ETO), a voting member of RTF, evaluated a smart thermostat energy efficiency program using RED and matching designs (Apex Analytics 2016).

Determining whether a common practice baseline approach provides appropriate savings estimates may depend on a jurisdiction's point of view and how these estimates are used within that jurisdiction. When used as part of a 5-year regional planning process, one point of view might emphasize the estimation of energy savings across five or ten program years. With this perspective, common practice baselines that are re-estimated periodically (as the RTF does) may reflect broad market changes over time. Common practice baselines change over time influenced, in part, by the on-going EE efforts over several years. An alternate view might be applied when looking at incremental resource investments. EE investments that offset other transmission and distribution (e.g., non-wires alternatives) and generation investments may focus on incremental energy savings which may be more appropriate. In this case, the fact that past EE programs may have changed the current EE baseline represents sunk costs and should not be considered economic assessments. If this is the case, only the savings that are incremental and attributable to that year's EE investments should be used. This illustrates how different

[69] To further illustrate, net savings as presented in the findings of EE evaluations are always presented as "net" of something; however, it may be gross savings net free-ridership, or it may be gross savings net free-ridership and spillover, or, in some cases, market effects may be included in the defined net savings estimates. Navigant (2013) found that most jurisdictions defined net savings as "gross savings adjusted only for free-ridership." (The review of net savings methodologies in Navigant [2013a] focused only on C&I programs. Of 38 C&I program evaluations reviewed, 28 estimated net savings as gross savings adjusted for free-ridership only. Three estimated net savings as gross adjusted for free-ridership plus participant spillover, and seven studies adjusted for free-ridership and both participant and nonparticipant spillover. None of the studies attempted to address market effects in addition to the spillover values.)

[70] One good example is the Bonneville Power Administration (2014) market characterization study of non-residential lighting in the northwest.

[71] RTF Guidelines (2015) state "The terms 'net' or 'gross' are intentionally not used to modify the term "savings," as they may conflict with the definition of baseline" in these guidelines.

[72] A presentation by Ms. Jennifer Light, RTF Chair, at the April 2017 Forum meeting presents context around the current practice baselines. https://nwcouncil.app.box.com/v/GuidelinesAprilRTFPres

jurisdictions may have different needs and uses for savings estimates, and how these can drive the approaches used (See NEEP, 2016).

Self-selection bias is a significant concern with common practice baselines. The average action taken in a current market may not be representative of those customers that chose to participate in a specific EE program. A common practice baseline will include a range of equipment with different levels of efficiency. An EE program that allows consumers to select themselves into the program may attract those consumers that comprise that portion of the common practice baseline who would have selected the high-end efficiency equipment. If an EE program attracted those consumers who were predisposed to install the high-efficiency equipment promoted by the program, application of a common practice baseline could overestimate net savings by not accounting for the unique characteristics of those customers. Additionally, to the extent that the program results in nonparticipant spillover, it is not clear how the common practice baseline approach would capture those savings. [73]

4.8.2 Constructing Common Practice Baselines

The theory underlying the definition and pros/cons of common practice baselines can be set out; however, there still is the task of developing these baselines. Developing and maintaining common practice baselines for all the individual measures included in a portfolio of residential and non-residential programs can be a daunting assignment. The RTF in the Northwest built up its library of measure protocols over several years.[74] In addition, the data and information needed for such these multiple baselines can be hard to develop.

SEE Action (2012b) indicates that appropriate common practice baselines can be estimated through surveys of participants and nonparticipants as well as analysis of market data. Discussions with the RTF indicate that they often scan websites of equipment providers to see what types of equipment are currently for sale online. In addition, there are also supporting studies characterizing the markets for energy-using equipment undertaken by other regional entities. A common practice baseline should be based on current sales of equipment and not on the stock of equipment installed. Sales of equipment will represent the current choices of equipment for customers. Sales data can be tough to come by and, even if available, may reflect only parts of the market. Access to sales data will vary by jurisdictions with those jurisdictions that have developed strong connections to equipment manufacturers, suppliers, and trade allies through long-standing EE programs likely to have better access.

When possible, the baseline should be substantiated by actual sales data from retailers and installers, not surveys and anecdotal information.[75] Considerations should include:

1. How much data is required to set a current practice baseline?
    a. Will additional market research and/or studies be required to set current practice baselines?
    b. How will minimum required confidence levels be determined?

[73] This will not be an issue in applications where market-wide sales data are available on standard and energy-efficient equipment, but these data are unavailable in most markets targeted by EE programs.
[74] The list of measures currently addressed by the RTF is available at https://rtf.nwcouncil.org/measures.
[75] Comment from Puget Sound Energy on the proposed baselines for non-residential lighting applications (2016).

2. Subcategories for common practice baselines—by business type, application, region? Regional variation was a significant issue for one IOU in the northwest as pricing and availability were viewed as varying across the service territory.

These considerations seem similar to those discussed in developing initial estimates of savings in Technical Resource Manuals (TRMs) that are developed by many states; however, some argue that common practice baselines with their implicit net savings construction might require a higher level of rigor.

Looking at how the RTF established its "current practice baselines" for two recent measure categories—residential lighting measures and non-residential lighting measures—can provide context for the approaches used for these applications. The non-residential lighting assessment involved the development of a matrix of fifteen applications/replacement types, and six candidate technologies. Not all technologies were appropriate for each application, and 43 efficiencies for incumbent technologies were developed. These were combined with market share estimates of sales to produce a common practice baseline for each of the fifteen applications. Documents available on the RTF website document this process.[76] An additional set of common practice baselines were developed for lighting controls. Also, a dual baseline was used to address installations that represented early replacement and accounted for the remaining useful life of the equipment that was replaced. The value developed for the remaining useful life was one of the more uncertain aspects of the baseline development. This energy savings protocol and baseline assessment represented one of the more complex efforts by the RTF and took more than a year to develop and approve.[77]

Another example of a common-practice type of baseline comes from California's efforts at developing Industry Standard Practices (ISPs).[78] This effort has a narrower focus addressing portable irrigation systems. This PG&E (2016) ISP study of portable irrigation piping systems had the following objectives:

- Evaluate the market trends.
- Determine the common industry practice materials used.
- Understand the barriers for adopting portable PVC (polyvinyl chloride) systems.
- Provide information and guidelines for California utility program developers and stakeholders to consider while developing and managing custom and/or deemed projects.

The data sources and methods used in this ISP included:

[76] The final RTF baseline recommendation for non-residential lighting can be found at: https://rtf.nwcouncil.org/meeting/rtf-meeting-december-6-2016. References to the process used and information from earlier efforts can be found at this site.
[77] Residential lighting energy savings protocols and baseline assessment was another complex effort undertaken. The process, data sources and results of this effort can be found on the RFT website at: https://rtf.nwcouncil.org/meeting/rtf-meeting-march-21-2017 and the supporting baseline presentation at https://nwcouncil.app.box.com/v/20170321ResLightPres.
[78] This example was provided by Dr. Tengfang (Tim) Xu at PG&E Customer Energy Solutions. Revisions to the current approaches for constructing ISP baselines as well as the use of ISP baselines in evaluation including their relationship to net savings are currently being considered (communication with Dr. Xu at PG&E).

1. Development and administration of surveys to customers (growers) and vendors/designers
2. Contacts with customers through emails and phone calls to compliment the surveys
3. Literature research
4. Analysis of the results to determine the market trends and market saturation of the available systems/technologies
5. Highlight critical issues that need to be addressed in custom project development.

Another consideration is that the common practice baseline is essentially a snapshot in time. The common practice baseline will change over time and periodic updates will be needed.[79] The complexity of the update will depend on the program type. For essentially a one-technology program (for example, refrigerator recycling), the update may be straightforward. Updating common practice baselines for a large C&I custom program where many technologies and end uses are impacted may be more difficult. In such cases, it might be more cost effective to focus exclusively on measures that account for the greatest savings. The RTF addresses this by establishing sunset provisions for each energy savings protocol and baseline(s). The sunsets vary across measure categories and are based on how fast the market is perceived to be moving—a market that is evolving rapidly would have a shorter sunset period.

4.8.3 Common Practice Baselines—Summary and Conclusions

Each example of common practice-types of baselines is a best estimate and is subject to uncertainty and potential bias. This is also true with the application of other methods for estimating net savings. Discussions with the engineering teams that developed common practice baselines at the RTF indicated that a practical approach is needed that:

1. Makes use of the best available data for each measure effort
2. Requires the energy savings and baseline engineering team to be open to input from other parties and modifications to initial baseline proposals
3. Develops agreement on the data which helps set the rules for estimating savings and allows for appropriate planning and consistency over time.

Common practice baselines have pros and cons. The decision to use this approach for certain measure categories and programs will depend in part on the jurisdiction's view of the needs and uses of energy savings estimates.

Ridge et al. (2013) make the point that previous EE programs have affected current markets for EE equipment through spillover and market effects. This results in current common practice baselines that are more efficient than they would have been if these past EE programs were not offered. The effect of these past programs is to lower the annual energy use of the measures that constitute the current practice.

[79] This is no different than programs evaluated using more traditional methods. The fundamental question is, "What is the shelf life of any evaluation given that many things (e.g., program intervention strategies, technologies promoted, targeted customers, and local and regional economic conditions) can change that would affect the program's ability to deliver net savings?" That is, all evaluations are essentially a snapshot in time.

This argument seems to be partly analytical and partly a policy consideration. Ideally, past evaluations of EE programs should have included all the impacts attributable to the programs, but because spillover and market effects were generally omitted from past evaluations, they have not been counted. The annual energy savings resulting from the use of common practice is lower than it would have been if these past programs were not offered. From this perspective, the use of unadjusted current practice baselines as estimates of net savings seems to be an effort to make up for mistakes in past evaluations (that is, the omission of spillover and market effects that impact the overall market). As a result, a jurisdiction may view savings that accrue today from programs in previous years along with the savings from current programs together comprise a reasonable estimate of EE program impacts over the long term; and, that this best represents the estimate of the overall return on investments in EE.

Another view or position is that each EE program should be evaluated as an incremental investment (that is, a program implemented in 2017 should be evaluated against what is attributable to that investment only—all impacts from prior years' programs are essentially sunk costs and should not be considered). This is an example of where policy and analytic views of net savings estimation are linked.

The bottom line for assessing the common practice baseline approach is the same process that is used in all other methods: (1) understand the construction of the baseline used in the evaluation; and (2) analyze the implications of this baseline against an appropriate counterfactual scenario for that program. The potential uncertainty and magnitude of bias needs to be at least subjectively assessed. Based on this standard approach, decisions can be made about the estimation methods most appropriate for the evaluation of an EE program taking into account jurisdictional priorities and needs.

When an evaluator encounters a jurisdiction that is using a "current practice baseline" method and refers to these savings as net savings, the evaluator should proceed in an internally consistent manner.[80] For example, it is important that the evaluator explain what the utility/agency/regional body is calling gross savings and what, if any, adjustments have been made in the establishment of the baseline to produce a net savings value.

In summary, several jurisdictions are looking toward the use of common practice baselines in their EE evaluation guidelines. As with all methods, there are pros and cons (see Table 7). A

[80] Reviewers of this section have commented that the evaluator might conduct multiple current baseline studies, calculate *ex post* net savings, and calculate a net realization rate to test the robustness of the approach; however, the cost of the analyses becomes a factor. Analyzing the market and different baselines has been presented as useful for understanding EE programs. This view may be most appropriate for jurisdictions that have EE measure and equipment specific data. These data may be limited to certain types of programs, and require a commitment to gathering data at the measure level. Also, before taking this approach, the evaluator might want to make sure that self-selection, nonparticipant spillover, and market effects are not serious sources of bias. If serious bias is suspected, the evaluator could select the baseline from the multiple baseline approaches above as the one that produces the most conservative results; however, there may be little analytic support for this selection. Another suggestion advanced in this newly developed literature is to augment the results using a survey based self-report NTG ratio, but this seems to defeat the purpose of using the common practice baseline method as an *ex ante* method of producing net savings. It increases costs and brings in the issues involved in using appropriate survey methods, and it may thereby reduce some of the advantages claimed for the common practice baseline approach.

potential strength of the common practice baseline approach is its use in upstream and market transformation EE programs. It can be applied market-wide and, unlike randomized trials and quasi-experimental designs, it does not require participants to be identified if appropriate sales data are available. However, this method is susceptible to self-selection bias (that is, the average consumer may not be the type of consumer who participates in the program). It is not clear how this can be addressed, other than by conducting surveys to determine specific characteristics of purchasers of efficient equipment relative to the common practice baseline. However, this survey effort would negate the unique aspects claimed for the common practice baseline approach; i.e., specific consumers who have and have not purchased the high efficiency equipment would need to be identified. This makes this approach more similar to the survey method approaches discussed in Section 4.2.

Table 11. Common Practice Baseline Approach—Summary View of Pros and Cons

Pros	Can help to avoid double counting of free-ridership in circumstances where gross impacts incorporate some net savings factors Can be used in upstream and market transformation programs Can be applied market-wide.
Cons	Self-selection bias is not addressed and methods for addressing self-selection are not readily apparent Does not capture nonparticipant spillover Common practice baselines for measures and technologies will change over time and require updating Determining average market practice has accuracy challenges Approach has been applied in the Pacific Northwest, along with other net savings estimation methods, but is relatively new and still evolving as a general net savings estimation method.

4.9 Top-Down Evaluations (Macroconsumption Models)

Top-down evaluations use macrodata on energy consumption in a model that relates changes in energy consumption to a measure of EE effort (usually expressed as EE expenditures). Top-down evaluation produced macroconsumption metrics (MCMs) in two recent pilot applications in California (see Cadmus 2012a; Demand Research, LLC 2012). The broader literature refers to these as top-down methods, and the MCM notation adopted in the recent California pilot studies refers to the same set of methods and cites top-down studies as background for its pilot work.

To date, this method's application has been somewhat limited. Applications to utility level programs have been limited to pilot studies and the general applicability of these methods has not been demonstrated. Still, the top-down approaches have appeal because they directly address overall net savings. The dependent variable is overall energy use (often expressed as energy use per capita) and this method simply examines the change in energy use resulting from EE efforts. Thus, there is in principle no need to adjust for free-ridership and spillover, or even for market effects, in estimating overall net savings. In addition, the regression analyses provide confidence and precision levels around these estimates. However, there are challenges in estimating the relationship between EE efforts and changes in overall energy consumption, such as the size of the impact isolated by the model.

The development of a model that can measure a 1%–2% change in total energy use annually and is attributable to EE programs requires a reasonably sophisticated structure. For example, the model must have an appropriate lag structure because the impacts from one year's expenditures will occur over several years.[81] In addition, the number of observations and quality of data needed to identify a small effect can be challenging. The data platform needed to support this top-down or MCM model approach requires the following:

- A measure of EE expenditures (or another metric of EE effort for different cross-sections, such as utilities or program administrators)
- Many observations to identify the effects of EE over several years, taking into account the lag structure of EE impacts. As a result, most top-down studies include multi-utility or multi-state efforts that can provide a reasonably large number of cross-sectional areas for the analyses
- Matching demographic and macroeconomic data to utility service areas, or subareas of utilities that are used as observations in the analyses
- High-quality data about energy consumption for each cross-section analyzed.

Questions that evaluators should consider when deciding on the appropriateness or applicability of top-down models are:

- What information will be produced by these top-down models if they are successfully estimated, recognizing that many cross-sections with varying levels of EE investment are needed for estimation?
- How does this information compare to what is produced by other methods?

Top-down models may be useful for:

- Estimating overall average change in energy use from the EE programs for a region. A top-down model that provides a good fit, meets reasonable assumptions, and has acceptable levels of statistical significance can provide information on the average change in overall energy use (or energy use per capita) from overall EE efforts.
- Estimating regional environmental impacts. Aggregate models can be useful in assessing state and regional environmental impacts such as the impact on carbon emissions.
- Providing evidence of estimated energy-savings at a regional level. The model can confirm—at an aggregate level—whether the expected energy savings are actually reflected in the macroconsumption data.
- Estimating overall cost savings from EE programs. Top-down models can also be used to estimate an overall cost savings per kilowatt-hour saved and confirm the efficacy of the overall EE effort.

[81] BC Hydro (2012) demonstrates the importance of the relationship between current expenditures on EE and future savings. It also shows the importance of letting the data determine the most appropriate lag structure as opposed to implementing a fixed structure that acts as a constraint. The estimate of energy savings is influenced by the manner in which lagged effects are handled in the regression model.

Top-down models, however, cannot provide information about:

- Savings produced by specific measures or programs or the impact of an individual program year for the over portfolio
- Where to make additional investments in EE at the program or measure level
- How to improve existing programs
- How to use estimates of free-ridership and spillover to suggest program improvements
- Quality assurance/quality control processes needed for regulatory oversight.

The relative importance jurisdictions and stakeholders place on program-level versus aggregated information will influence decisions to implement these types of evaluation frameworks. Top-down approaches seem complementary to results produced by program-level evaluations; however, there may be concerns about using these methods to replace program-level evaluations. Some view the program-level research as essential in that it helps ensure that the right set of programs comprise the EE portfolio and it is useful in addressing program- and portfolio-specific questions about implementation. Top-down methods and program-level evaluation provide useful, but different, perspectives on the accomplishments of EE efforts.

Cadmus (2012a) reviewed a number of top-down studies that expressed energy consumption as a function of a metric meant to measure EE effort including:

- Parfomak and Lave (1996) used a panel dataset of 39 utilities from 1970 to 1993. The claimed savings by utilities for their C&I programs was used as a proxy for the level of EE effort. The regression analysis was similar to a realization rate regression analysis model, where the coefficient on the claimed utility savings indicated what fraction of those savings could be found in the data. The authors estimated the realization rate for the utility's claimed savings at 99%.
- Auffhammer et al. (2008)—working with data developed by Loughran and Kulick (2004)—used what has become the more traditional formulation. Here, EE effort was expressed in the econometric model as program expenditures reported to the U.S. Energy Information Administration. The authors found that average utility reported savings (2%–3%) fell within the 95% confidence interval for estimated savings. The cost of saved energy was approximately $0.06/kWh.
- Arimura et al. (2011) also used the Energy Information Administration data on program expenditures across 307 U.S. utilities to examine the impacts of EE investments on overall energy consumption.[82] The authors used utility Energy Information Administration data from 1989 to 2006 to determine electricity savings of 1.8% annually and estimated the cost of saved energy at approximately $0.05/kWh.

[82] Arimura et al. (2011) also advance the state of the practice by modeling energy prices and utility EE program expenditures as endogenous and allowing consumption to depend on program expenditures in a flexible way. The literature on top-down models represents sophisticated applications of econometric methods. Problems of endogeneity and autocorrelation with flexible lag structures have become common issues that are addressed by these models.

The California Pilot Project on top-down methods involved two efforts, Cadmus (2012a) and Demand Research, LLC (2012).

In addition to these studies presented below as case studies, the Massachusetts Program Administrators have been piloting top-down studies since 2015. As part of this effort, the evaluation team recently completed a more extensive literature review of top down methodologies applied in studies in both the energy and non-energy sectors. They also conducted in-depth interview with nationally recognized experts in the field of econometrics, macroeconomics, and top down modeling. Some of the methodological considerations examined in the literature review and industry expert interviews included the best approach for handling seasonal weather variations, how top down models should account for the cumulative effects of energy efficiency programs over time, how models should determine the impacts on energy usage for the recession years (2007-2009), the types of fixed effects terms that should be included in a model, as well as considerations from determining error bounds form the model results.

Based on the results of the literature review and expert interviews, the evaluation team provided an assessment of the advantages and disadvantages of different theoretical methodologies and approaches, and provided recommendations for enhancements to future top-down efforts in a February 16, 2017 report to the Massachusetts Program Administrators.[83]

4.9.1 Case Studies on Top-Down Approaches

Example 1: Cadmus California Top-Down Pilot Study

Cadmus used expenditures on EE programs as the level of EE effort in its models. The models were estimated at the utility level for residential and nonresidential energy savings. Cadmus worked with data at the utility level using information from the three investor-owned utilities (IOUs) and from large public utilities in California such as Los Angeles Department of Water and Power and the Sacramento Municipal Utility District. Data were also collected from some small public utilities, but were generally inconsistent.

Several models estimated the relationship between utility energy consumption for residential and nonresidential customer segments and EE expenditures.[84] Overall, it was difficult to obtain significant results across the models. The best model produced significant coefficients on the EE expenditures variable using only data from the three IOUs. To demonstrate the information that can be produced by top-down models, Cadmus developed estimates of savings from EE efforts over a 6-year period and calculated the cost of energy saved. Savings from EE spending from 2005 to 2010 were estimated at 8%, and the cost per kilowatt-hour saved was estimated at \$0.05. The results of the Cadmus study indicated savings were within 10% of the net savings reported by California IOUs for the 2006 to 2008 program cycle. The estimates of energy savings and cost per kilowatt-hour saved had large confidence intervals: ±66% on the energy savings

[83] Tetra Tech, NMR Group, and DNV GL (2017). Top-Down Modeling Extended Methods Review. Prepared for the Massachusetts Program Administrators, February 16, 2017.

[84] Cadmus (2012a) did not try to estimate separate models for commercial and industrial consumers because the time series was inconsistent. In some years, commercial sector consumption would increase and industrial consumption would decrease by approximately the same amount. This suggested that there was some switching in the definition on the commercial and industrial rate classes. As a result, the two classes were modeled together.

estimate and more than ±100% on cost per kilowatt-hour saved. The 48 observations in the top-down IOU model resulted in lower precision than studies with much larger sample sizes.

Cadmus did consider disaggregating the data beyond the IOU level to gain more cross-sections for the analysis; however, there was concern about the ability to allocate EE program expenditures to smaller geographic areas. One specific concern was the savings from CFLs. More than 50% of the expected savings were from CFLs and these sales were tracked at point of sale instead of the location where they were used, making it difficult to align the energy consumption and the impact of EE expenditures for smaller geographic areas.

Example 2: Demand Research, LLC, California Top-Down Pilot Study

Demand Research (2012) developed an MCM model working with California utilities and program contractors that disaggregated residential energy use and estimates of residential sector EE efforts into a database of cross-sectional observations at the census tract level. C&I sector energy use and metrics for EE efforts were disaggregated down to the county level. Instead of using energy expenditures, the Demand Research, LLC, study used the utilities' *ex ante* estimates of energy saved by census tract as the metric of residential EE effort. Parfomak and Lave (1996) used a similar approach. For the C&I sectors, county-level data were developed. The independent variable for the EE level of effort in the commercial sector model was a metric related to incentives paid; however, *ex ante* energy savings was used as the metric for EE effort by county for the industrial sector.[85, 86]

The findings from the Demand Research, LLC, study were:

- The residential models estimated by Demand Research, LLC, (2012) showed that higher levels of the EE effort variable resulted in reduced energy use with statistically significant estimates at a 95% confidence interval.
- The commercial sector model produced the expected sign on the EE effort variable, but the results were not statistically significant.
- The industrial sector model did produce statistically significant results for the EE effort variable.
- The residential and C&I sector models produced statewide savings estimates of 7.3% for the 5-year period from 2006 to 2010.

[85] Different metrics for EE level of effort were used in the C&I sector model because the method selected to address endogeneity in the commercial sector model ensured that the EE level of effort variables uncorrelated with the error term.

[86] Considerable work went into creating the census tract databases for the residential model and the county level databases used in the commercial and industrial models. The details can be found in the full study, but as an overview of the effort -- key energy consumption and program tracking data by fuel and segment were inspected prior to modeling for missing values, seemingly erroneous data or outliers, and high and low-end values that might skew the sample statistics or suggest multimodal distributions. Other adjustments to the datasets were made, including the use of a "restricted" commercial sector dataset that included only counties with high *ex ante* energy savings values in this pilot test. Dropping sites from statistical analyses that likely provide no information because the expected savings from those sites are so small is not uncommon. The usual justification is that the total savings number is not likely to be influenced by their exclusion because the expected savings were so small.

- The relative precision for the aggregate savings estimate was ±31% (or a 90% confidence interval of 5.0%–9.5%).
- The estimated statewide savings of 7.3% exceeded the utility *ex ante* estimates of 4.8%.

The aggregate statewide estimate of energy savings across all three sectors was forecasted with reasonable confidence and precision. Looking at the results at one level of disaggregation lower (at the sector level results) shows a high degree of variability. For example:

- The estimated industrial energy savings (all three utilities combined) were about 745% higher than the utilities' *ex ante* values (Demand Research, LLC 2012, p. 36).
- The commercial sector kilowatt-hour savings estimates (all three IOUs combined) were about 27% lower than the utilities' *ex ante* estimates.
- The residential sector savings estimates from the MCM model for Pacific Gas & Electric and San Diego Gas & Electric (Southern California Edison was not estimated) were substantially higher than the utilities' *ex ante* values.

When these sector-level results are aggregated up to a statewide number, the wide discrepancies at the sector level tend to offset each other. It is important to recognize that this was a pilot effort and views will differ about the overall robustness of findings at the sector and statewide levels.

4.9.2 Developing Top-Down Models

Cadmus (2012a) and Demand Research, LLC, (2012) took different paths to developing a top-down MCM model for this California Pilot Study. Both study teams concluded that the work to date indicated this was a potentially useful research path for developing statewide estimates of energy savings attributable to EE policies. In its study report, Cadmus discussed the potential applications of these methods:

- Top-down macroconsumption methods could yield inexpensive[87] estimates of energy savings from utility EE programs and building codes at an aggregate level.
- These methods are attractive because it is possible to produce confidence and precision levels for the net energy savings estimates, which is not as easily accomplished in bottom-up evaluation studies. [88]

[87] Both pilot studies ran into data problems that would have to be overcome in future work and could be costly to address. If the alternative were to build up statewide estimates by doing measure-specific engineering analyses, this aggregate Top-Down approach might be less expensive; however, bottom-up methods performed cost effectively are probably needed for program support, design, and verification of savings at the program level. The issue is whether the incremental information provided by these aggregate studies has a value greater than its cost. That may vary by jurisdiction.

[88] This is a conclusion from the Cadmus (2012a) top-down applications; however, bottom-up approaches also routinely calculate confidence and precision levels for program and portfolio estimates of net savings. The advantage with the top-down approach might be that the confidence and precision levels can be calculated more easily at the aggregate level, because different values for confidence and precision across programs do not have to be combined using assumptions about the covariance across the different distributions from which these values are calculated for each program.

- Top-down studies can be used to verify statewide EE program savings estimates based on bottom-up evaluation by looking at aggregate energy consumption data.
- These methods can be useful in tracking a state's progress in reducing greenhouse gas emissions and developing forecasts of energy savings from future program spending at an aggregate level.

Next steps that might provide additional insights into this top-down application are to: (1) replicate the results of Cadmus and Demand Research, LLC using the datasets already developed; and (2) continue improving the data platform used for these analyses—both studies contained recommendations for improving the data. Violette et al. (2012) discuss the importance of the data platform on which these top-down models are estimated. Other considerations pertain to the sensitivity of the results to model specification (that is, the robustness of the results under a designed set of alternative specifications that are also consistent with the theory and appropriate econometric methods).[89]

Top-down studies cannot entirely replace bottom-up studies (see Table 12 for pros and cons of these methods). As discussed earlier, there is likely a need to have program-level (and some measure-level) assessments to ensure that a program's design will result in a program meeting its specified targets. Moreover, top-down studies are subject to a range of methodological uncertainties not fully captured by the measured precision, just as the bottom-up estimates are. Evaluators should ask, "Does the incremental value of the information produced by the top-down methods exceed the cost of the work?" At the national level, data from an adequate number of cross-sectional observations are more easily available. For state-level studies, more work will be involved in setting up the databases and disaggregating the data into the number of needed cross-sections, which may introduce some error into these observations.[90]

[89] This sensitivity analysis might examine the stability of the estimates under alternative functional forms, inclusion of one or two variables, testing of interaction terms, and tests on subsets of the data.
[90] Violette and Provencher (2012) discuss attenuation bias where the coefficients on independent variable can be biased toward zero due to errors in the measurement of variables. A similar effect is shown in Ridge (1997).

Table 12. Top-Down Evaluations (Macroeconomic Models)—Summary View of Pros and Cons

Pros	Estimates net effects of all programs cumulatively No need to adjust for free-ridership, spillover, or market effects at the aggregate level.
Cons	Methods are not fully developed at the state or regional levels Relies on high-quality energy consumption data and on data regarding EE efforts within each cross-section analyzed Subject to bias and uncertainty due to self-selection, cross-unit spillover, data limitations, and model specification uncertainty Cannot provide savings at the measure, technology, or program level Does not provide information on how to improve program design and implementation processes.

5 Conclusions and Recommendations

A central theme in this chapter is that all decisions have an implicit counterfactual scenario—what would have happened if the decision had not been made. In the context of EE programs, net savings are energy use with the program as compared to a counterfactual that is meant to represent what energy use would have been without the program investments. This chapter does not prescribe specific methods for determining net savings, but rather it presents approaches for assessing attribution and the net impacts of EE programs and discusses the issues affecting the choice of a net savings approach within an evaluation context.

5.1 A Layered Evaluation Approach

It is important that the selected approach be appropriate for the intended audience and present analyses supported by evidence. A well-executed statistical analysis may be a central piece of the evaluation, but it still may not be persuasive to many decision-makers and stakeholders on its own. All approaches should be supported by a narrative discussing why a specific approach was taken, the appropriate interpretation of the findings, and the context for identifying net savings (see historical tracing above). The narrative and analysis should also recognize and indicate the uncertainty in net savings determination. Developing an appropriate narrative often leads to the application of layered methods of analyses.

Studies examining net savings from EE programs may contain both sophisticated quantitative analyses as well as intuitive analyses that show savings that are attributable to the program exist. A compelling part of the narrative can be a simple case study of one or two market participants. A case study can show with a very high degree of internal validity that net savings were obtained, and/or provide examples of NTG factors including free-ridership, spillover, and market effects. An intuitive case study often is a useful first step in a two-part analysis framework to address estimates of net savings. For example:

- **Part 1:** Establish the existence of the effect, possibly using a case study approach. This can include establishing the existence of savings that are attributable to the program. If the focus of the research is on estimating free-ridership or spillover, the first step can involve establishing the existence of these effects. Once existence of an effect is established, the magnitude of the effect needs to be determined. This can be easier when the audience is convinced that the effect exists (i.e., the effect is nonzero), and the logic behind the attribution of the effect is set out.
- **Part 2:** This involves the extrapolation of the findings of the case studies to the more general participant population. Once the logic of the case studies is established, it is often possible to define and apply a statistical model consistent with this logic, or to develop an alternative approach to extrapolate the effect. This approach could include any of the methods discussed in this chapter—survey methods, common practice baselines, market data analyses and comparisons, structured expert surveys, or historical tracing to examine the influence of a program over time.

The framework above for analyzing net savings can be extended to three steps:

1. Perform an initial high internal validity case study to prove the existence of effects.

2. Establish an estimate range. In other words, determine a reasonable lower bound for the impacts and the highest reasonable bound from the evaluation analyses. This provides information about the importance of the studied effect and whether it is a part of net savings or an NTG factor (free-ridership, spillover, or market effect).
3. Perform analyses using the methods presented in this chapter to develop the best estimate of impacts within the established range.[91]

5.2 Selecting the Primary Estimation Method

The selection of appropriate net savings analysis methods will depend in part on the questions that need to be answered by a net savings study. Research issues that have implications for the net savings approach include:

- **RCTs and quasi-experimental designs** employing DiD and regression methods along with RDD and RED designs (discussed in Sections 4.1 and 4.2 of this chapter). These approaches produce estimates of net savings that address free-ridership and participant spillover. Nonparticipant spillover is not directly addressed but can be addressed through surveys of nonparticipants and market effects studies with trade allies.
- **Survey methods** can be used to adjust engineering-based gross savings estimates for free-ridership and participant spillover (discussed in Section 4.3). Nonparticipant spillover can be addressed through surveys of nonparticipants and market effects studies using trade allies.
- **Broader-based methods such as market sales, structured judgment, and historical tracing analyses** can all be used to provide program-specific net savings estimates and address spillover and market effects (discussed in Sections 4.4, 4.5, and 4.7).
- **Deemed or stipulated methods** can be set at the program level (discussed in Section 4.6); however, the applicability from one jurisdiction to another should be considered.
- **Common practice baseline methods** can produce estimates by developing baselines on a program basis (discussed in Section 4.8). This approach may not fully address free-ridership or participant spillover, because it does not account for self-selection bias. Also, it does not directly address nonparticipant spillover. However, as previously noted, nonparticipant spillover can be addressed through surveys of nonparticipants and market effects studies with trade allies. Common practice baseline methods might be viewed as a compromise that balances out over- and underestimated NTG factors in the net savings estimate.
- **Top-down analyses** use aggregate data that represent the overall level of EE effort across all programs, but cannot isolate the effects of a single program or measure (discussed in

[91] In a survey setting, this approach can help the survey respondent consider first the behavior that might result in lower, and then the higher impacts that might have been achieved if the program had not existed. The thought process developed by this three-step construct can help survey respondents produce better estimates of their most likely behavior by first thinking through a construct where the respondent is first asked about factors that would result in a low-range value and then factors that would result in a high-range value.

Section 4.9). Top-down models conceptually address all of the NTG factors—free-ridership, spillover, and market effects.

How can estimates of net savings on a program basis be combined with information about program implementation effectiveness? Approaches that provide estimates of net savings but also include elements that involve gathering information directly from participants, nonparticipants, and trade allies can be useful for improving program performance. For example, some programs are designed to minimize free-ridership to improve overall resource effectiveness and others focus on expanding the magnitude of spillover and market effects. For these programs, specific estimates of free-ridership, spillover, and market effects—particularly if they are provided over a longer time period (every 2 years)—can be used to assess overall program effectiveness.

Can evaluators estimate aggregate net savings from a portfolio of programs? All the estimation approaches presented here, except the top-down analyses, can produce program-specific estimates that evaluators can aggregate up to the portfolio level. Top-down methods are designed to work with aggregate data, particularly at the regional level.

Other factors that influence the selection of appropriate methods will vary by program type, delivery, sector, and maturity. A recent free-ridership and spillover methodology study for the Massachusetts Program Administrators describes the key elements evaluators should consider when choosing a method (Tetra Tech et al. 2011). This study addressed the following factors:

- **Availability of market sales data with a meaningful comparison group**. If market sales data are available on the total sales of both efficient and standard equipment over time, these data are available for the program area, and there is an appropriate comparison area for the appropriate time, total program effects may be estimated based on these data.

  The ideal strategy is to compare the magnitude of the change in sales of energy-efficient equipment relative to the sales of standard equipment in the program area and the comparison area. However, the program tends to produce systematic differences between the program and comparison areas. Therefore, where a program has been operating for a long period of time, it is very difficult to find a comparable comparison area.

- **Homogeneity of the measure and the consumers**. RCTs and quasi-experimental designs work best when there are many similar consumer types and measures. Large custom programs are likely to have fewer projects, so a few (or even one) very large project(s) can have a significant influence on free-ridership or spillover. Therefore, the evaluator should use multiple approaches that allow for a greater focus on the consumers that drive the overall impacts to confirm the findings for that program. Methods based on market data or samples of consumers who are making similar purchase decisions may not apply to programs with custom measures.

- **Likelihood of substantial upstream effects unknown to end-use participants**.[92] If there is a reasonable likelihood of substantial upstream effects that an end-use participant would not know about, then conducting an evaluation by using participating end-user surveys alone will tend to understate the effect of the program (even if consumers answer accurately from their perspectives). These situations require either information for the market as a whole (if the market sales-based approach is viable) or a combination of participant end-user and vendor surveys.
- **Cost/value tradeoffs**. Some methods that provide more credible results are costlier. This cost may be justified for program components that are important to the portfolio, but not for all components. Importance to the portfolio is typically related to the level of spending or savings associated with a program component. However, a component's importance can also depend on future program plans or other "visibility" factors. The systematic assessment of the value of information gained by net savings estimation approaches s compared to the cost of the research is needed to better balance the requests to meet confidence and precision levels for estimates. A target of 90% confidence at ±10% precision simply may not be reasonable for all but the largest programs in a portfolio. This systematic approach can examine the impacts on ratepayers from incorrectly attributing savings to a program. If it is a small program, the impacts on ratepayers will be small as measured with 90% confidence and 15% or 20% precision using a one-tailed test. This can substantively reduce evaluation costs with little impact on the overall equity tradeoffs between ratepayers and utilities.
- **Data quality**. Data quality is a critical factor for all methods. Typical examples of potential limitations to good data quality are: (1) insufficient information in program tracking databases; (2) lack of clear definitions of what is contained in tracking systems (that is, a data dictionary); (3) limitations on the availability of nonparticipant data (including billing data); (4) insufficient number of years of available billing data for participants; and (5) limitations on the availability of market sales data.

5.3 Methods Applicable for Different Conditions

Table 13 lists methods that are suitable for programs with particular features (based on Tetra Tech et al. [2011]). Programs operate in a context and choosing the appropriate evaluation methods requires balancing the advantages and disadvantages of each method. Thus, this table does not list recommendations for a preferred method for a given situation. Rather, it indicates which of the available methods are applicable to programs with specific features. The scales (i.e., low to high) represented in the table for typical cost and complexity are meant to provide an indication of applicability and cost or complexity relative to other methods in Table 13.

[92] For example, the participating customer may not know that the program influence has changed what options are available, lowered the price of the efficient options, and/or increased the sales staff's knowledge and interest in promoting the efficient option.

Table 13. Summary of Methods Applicable to Different Conditions

Net Savings Method	Surveyed Group	Applicability				Typical Cost or Complexity	Special Requirements
		Custom Measures	Measures With Few, Diverse Participants	Large Numbers of Similar Participants	Measures With Substantial Upstream Influence Invisible to Consumers		
RCTs using DiD	None necessary, but could be conducted to help validate the baseline as an appropriate counterfactual scenario	Poor	Poor	Good	Poor	Low	Random assignment of participants and controls
Quasi-experimental design	None necessary but could be conducted to validate or develop better baselines	Poor	Poor	Good	Poor	Low	Matched nonparticipant comparison group
Regression models—Billing data analyses with control variables and Linear Fixed Effects Regression (LFER)	Participating consumers and comparison group consumers	Poor	Poor	Good if there is a valid comparison group	Good if there is a valid comparison group	Low	Need control variables that influence energy use across participants and nonparticipants
Survey based—participants, nonparticipants, and market actors	Participating end users	Good	Good	Good	Poor unless combined with retailer or contractor surveys	Medium	Counterfactual baseline based on survey responses
	Participating and nonparticipating end users	Poor	Poor	Good	Poor unless combined with retailer or contractor surveys	Medium-High	Nonparticipants must be representative of participants
	Retail store managers and contractors	Good	Good	Medium	Good	Medium	
Survey	Retail store managers	Poor	Poor	Good	Good	Low	

Net Savings Method	Surveyed Group	Applicability: Custom Measures	Applicability: Measures With Few, Diverse Participants	Applicability: Large Numbers of Similar Participants	Applicability: Measures With Substantial Upstream Influence Invisible to Consumers	Typical Cost or Complexity	Special Requirements
based - qualitative sales and counterfactual scenario	and contractors						
Structured expert judgment	Experts	Depends on quality of input methods				Low	
Market sales data (cross-sectional studies)	None	Poor	Poor	Good	Good	Low if data are available; high or not possible if data must be developed	Defined market segment
	Manufacturers and regional buyers and distributors	Poor	Poor	Good	Good	Low	
	Retail store managers and contractors	Good	Good	Medium	Good	Medium	
Common practice baseline	Participating and Nonparticipating end-user surveys or market sales data are used	Poor	Poor	Good	Good	Medium to high	Defined market segment
Top-down methods for regional application	None	Requires data on aggregate energy consumption and information on EE effort (expenditures or related program variable) for a large number of cross-sectional observations over a period of time				Depends on the cost of compiling the initial dataset	Aggregate data available on geographic cross-sections

5.4 Planning Net Savings Evaluations—Issues to Be Considered

Evaluation planners should consider several practical issues when planning a net savings evaluation. These include the use of the information, maturity of the program, timing of the study, frequency of net savings estimation, and whether to use multiple approaches. The following bullets summarize these issues:

- **Use of the information**. It is important to consider how the results of the net savings evaluation will be used and the audience for which the evaluation is intended. This can include shareholder incentives, resource plans, program design, and environmental targets (for example, carbon emissions), among other policy goals.[93] The Gross and Net Savings Decision Making Framework and Template (NEEP 2016) provides a standalone template to guide and document key elements that should be considered when making policy decisions in which gross and/or net savings from energy efficiency programs play a role.
- **Maturity of the program**. Almost all programs are assumed to have some free-ridership. The conventional wisdom is that as the program matures (all else equal), observed free-ridership will increase during the study period, but so will spillover and market effects. As a result, it becomes important to test for spillover and market effects as a program matures.
- **Timing of data collection**. To estimate free-ridership, the data should be collected as soon as possible after program participation. This timely measurement minimizes recall bias (Baumgartner 2013), provides apt feedback on program design, and reduces the possibility that the key decision-maker or market actor is no longer available. However, if the objective is to estimate spillover, the ideal time to collect data is at least 1–2 years after program participation, as this allows sufficient time for spillover to occur. Finally, if the objective is to estimate market effects, regular data collection over a period of time is required.
- **Frequency of net savings estimation**. The frequency of net savings or NTG analyses depends on the use of the information. If it is a component of financial incentives for a program administrator, evaluators may need to conduct these studies more frequently. Usually, there is no need to perform detailed net savings studies more than every other year. But, it also depends on the methods used. A statistical analysis of a residential behavioral program can be estimated every year, because persistence is an important issue and study costs are low. NEEP recommends that net savings estimates be made every 2–5 years (Titus and Michals 2008) because several factors can cause estimates of net savings to change over time.
- **Triangulation of NTG approaches**. Using data from multiple sources limits the effects of self-report bias and measurement error (Baumgartner 2013). Using an in-depth methodology with multiple sources also allows evaluators to weight the value of

[93] For example, NEEP (2012) showed that “compared to New England and New York, states in the Mid-Atlantic more commonly use evaluated gross savings for utility regulatory compliance and net savings for program planning and measurement of cost effectiveness. In contrast, New England and New York are more likely to use evaluated net savings; in doing so, they apply NTG values prospectively rather than retrospectively.”

responses from different decision-makers (Megdal et al. 2009). Other data sources often used are: (1) interviews with key decision-makers at the site; (2) project file reviews or project analysis that looks at barriers to project installation, how the project addressed those barriers, and documentation on the participant's decision to go forward with the project; and (3) market data collection, which might include analyses of market sales and shipping data and surveys of market actors (GDS Associates, Inc. et al. 2010; SEE Action 2012b). A recent study conducted for the Massachusetts Program Administrators presents a general approach that can be used by others as they seek to triangulate and integrate the results of two or more net savings studies.[94] The general approach organizes the results from each study in a table that shows findings for each net savings component as well as the qualities or key considerations of each study's results. This approach provides transparency in the factors driving the final net savings estimate.

- **Some evaluation issues are best addressed prior to rolling out a new or revised EE program**. Program design personnel and evaluators should work together in advance of implementing a program design that includes random assignment to discuss the data needed for evaluation that must be collected as part of program implementation.

5.5 Trends and Recommendations in Estimating Net Savings

As discussed in Section 5.4, the choice of approach for estimating net savings will vary depending on the questions asked, the characteristics of the program(s) evaluated, and the ultimate use of the data. However, there are trends in the application of methods:

- The expanded use of informational and behavioral EE programs is leading to a greater use of RCTs and quasi-experimental designs that employ some form of randomization (RDD or RED) to help address self-selection.
- The complexity of programs and the need for assessing market effects is leading to a greater use of informed expert panels and Delphi-types of analyses.
- The need to examine trends in program performance over time and impacts on markets over time is resulting in long-term planning for net savings and NTG factor analyses (for example, regular studies conducted with panel data).
- Net savings studies are increasingly embedded in survey analyses that are also designed to gather information about program implementation effectiveness.
- The value of information from net savings studies is being considered in a more structured manner to help manage evaluation costs (see NEEP 2016). Achieving 90% confidence and 10% precision may be important for a very large EE program, but for a program that is one tenth of the size of the largest program, precision levels are being generated that represent only 1% of the large program. Also, one-tailed tests should be considered, because for some applications, it may be more important to attain a threshold level of net savings with a certain level of confidence than it is to bound the savings estimate both above and below using a two-tailed test. A one-tailed targeted precision level still allows for the calculation of the upper end to the confidence interval Violette

[94] Tetra Tech; NMR Group; DNV GL (2017). *Net-to-Gross Methodology Research* (TXC08)

and Rogers (2012), and there is value to knowing if there was a high likelihood that the target was exceeded by a given amount. The appropriate level of confidence and precision targets are now often reviewed by EE program administrators and regulators to provide fair attribution estimates that minimize risks to ratepayers and to utilities receiving incentives. Navigant (2013a) discusses a loss function approach for assessing the value of information from net savings studies; and information on sampling and the tradeoffs between confidence and precision for EE evaluation can be found in Violette and Rogers (2012) and Khawaja et al. (2013).

It has always been important to consider evaluation options before implementing an EE program or portfolio of programs. However, the importance of planning the types of net savings studies that are needed and the frequency of this measurement prior to program implementation are becoming critically important. Net savings studies embedded in experimental designs that are established prior to consumers becoming program participants allow for:

- The consideration of randomized designs
- The development of the data platform for estimating consumption-based models (including top-down models)
- The collection of information needed for well-run structured expert panel studies.

In conclusion, net savings methodologies continue to evolve and improve over time. No single methodology is appropriate for all programs or measures, and a single methodology is often not the best choice for estimating program or measure net savings. In the end, jurisdictions should design evaluation plans to assess net savings in conjunction with the key stakeholders considering:

- The appropriate schedule for the evaluation effort over time, taking into account the expected value of the information produced versus the cost of the research effort
- Program design and maturity
- The contribution of the program to overall portfolio savings (past, current, planned)
- The evaluation budget, objectives, and value
- Observations and lessons learned from other jurisdictions.

Finally, adequately documenting the methods used and effectively communicating the results of any net savings study are important. The beginning of this chapter presents a framework for persuasive communication.

6 References

AEP (2012). Appendix H, Evaluation of Home Energy Reports, prepared by Navigant. http://www.oracle.com/us/industries/utilities/home-energy-reports-program-3631943.pdf

Agnew, K.; Goldberg, M. (2017). *Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol: Methods for Determining Energy Efficiency Savings for Specific Measure*s. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68564. http://www.nrel.gov/docs/fy17osti/68564.pdf

Angrist, J.D.; Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Apex Analytics (2016). Energy Trust of Oregon Smart Thermostat Pilot Evaluation. Prepared for Energy Trust of Oregon. March 1. http://assets.energytrust.org/api/assets/reports/Smart_Thermostat_Pilot_Evaluation-Final_wSR.pdf

Arimura, T.H.; Li, S.; Newell, R.G.; Palmer, K. (2011). *Cost Effectiveness of Electricity Energy Efficiency Programs.* Resources for the Future Discussion Paper 09-48-Rev.

Auffhammer, M.; Blumstein, C.; Fowlie, M. (2008). "Demand-Side Management and Energy Efficiency Revisited." *The Energy Journal* 29(3): 91–103.

Baumgartner, R. (2013). "Survey Design and Implementation Cross-Cutting Protocols for Estimating Gross Savings." Chapter 12 in *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*. www.nrel.gov/docs/fy13osti/53827.pdf.

BC Hydro (2012). *Review of a Top Down Evaluation Study: Rivers & Jaccard*. Prepared for BC Hydro, by Navigant Consulting, Inc., April.

Bodmann, S. (2013). Controlling for Program Participation Self-Selection Bias. Paper and Presentation at the International Energy Program Evaluation Conference.

Bonneville Power Administration (2014). *Northwest Nonresidential Lighting Market Characterization 2010-2012*. Navigant Consulting, Inc. Cadeo Group, https://www.bpa.gov/EE/Utility/research-archive/Documents/Northwest_NonRes_Lighting_Market_Characterization.pdf

Bradlow, E. (1998). "Encouragement Designs: An Approach to Self-Selected Samples in an Experimental Design." *Marketing Letters,* 9(4).

Cadmus (2012a). *CPUC Macro Consumption Metric Pilot Study: Final Report*. Prepared for the California Public Utilities Commission, October.

Cadmus (2012b). *Efficiency Maine Trust Residential Lighting Program Evaluation: Final Report.* Prepared for the Efficiency Maine Trust. www.efficiencymaine.com/docs/Efficiency-Maine-Residential-Lighting-Program-Final-Report_FINAL.pdf

Cadmus; Navigant (2012). *New York Energy $mart[sm] Products Program Market Characterization and Assessment Evaluation: Final Report.* Prepared for The New York State Energy Research and Development Authority, Victoria Engel-Fowles Project Manager, Project Number 9875, February. See bullet six study in the link: https://library.cee1.org/system/files/library/10360/CEE_Eval_New_York_Energy_Smart_Products_Program_Market_Characterization_and_Assessment_Evaluation_Feb2012.pdf

Cadmus; Navigant Consulting; Opinion Dynamics Corporation (2012). *2012 Residential Heating, Water Heating, and Cooling Equipment Evaluation: Net-to-Gross, Market Effects, and Equipment Replacement Timing*. Prepared for the Electric and Gas Program Administrators of Massachusetts.

Castor, S. (2012). *Fast Feedback Results.* 2011 Final Report prepared for Energy Trust of Oregon. https://www.energytrust.org/wp-content/uploads/2016/12/Fast_Feedback_-_20110.pdf

Commonwealth Edison (2012). *Energy Efficiency / Demand Response Plan: Plan Year 3 (6/1/2010-5/31/2011) Evaluation Report: Home Energy Reports*. Prepared by Navigant Consulting. www.icc.illinois.gov/downloads/public/edocket/323839.pdf

Cook, T.; Scriven, M.; Coryn, C.L.; Evergreen, S.D.H. (2010). "Contemporary Thinking About Causation in Evaluation." *American Journal of Evaluation* 31:105. http://aje.sagepub.com/content/31/1/105

Demand Research, LLC (2012). *Macro Consumption Metrics Pilot Study: Final Repor*t. Prepared for: California Public Utilities Commission Energy Division, November.

Diamond, A.; Haninmueller, J. (2007). *The Encouragement Design for Program Evaluation*. Harvard University and International Finance Corporation.

DNV GL, ERS (2017). *Massachusetts Commercial/Industrial Baseline Framework.* Prepared for the Massachusetts Program Administrators and Energy Efficiency Advisory Council. http://ma-eeac.org/wordpress/wp-content/uploads/MA-Commercial-and-Industrial-Baseline-Framework-1.pdf

DNV GL (2017). *Impact Evaluation of 2015 Upstream and Residential Downstream Lighting Programs*. Prepared for the California Public Utilities Commission. CALMAC Study ID CPU0152.01.

http://www.calmac.org/publications/2015_LTG4_Impact_Evaluation_Report-FINALES.pdf

Dubin, J.; McFadden, D. (1984). "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption." *Econometrica* 52(2):345–362

Duflo, E.; Glennerster, R.; Kremer, M. (2007). *Using Randomization in Development Economics Research: A Toolkit.* Centre for Economic Policy Research, Discussion Paper 6059.

Eto, J. (1988). "On Using Degree-days to Account for the Effects of Weather on Annual Energy Use in Office Buildings." *Energy and Building* 12, 113–127.

Eto, J.; Prahl, R.; Schlegal J. (1996). *A Scoping Study on Energy-efficiency Market Transformation by California Utility DSM Programs*. Lawrence Berkeley National Laboratory. http://emp.lbl.gov/sites/all/files/lbnl-39058.pdf .

Fagan, J.; Messenger, M.; Rufo, M.; Lai, P. (2009). "A Meta-Analysis of Net to Gross Estimates in California." Paper presented at the 2009 AESP conference.

Feng, W.; Jun, Y.; Xu, R. (2006). *A Method/Macro Based on Propensity Score and Mahalanobis Distance to Reduce Bias in Treatment Comparison in Observational Study*. Public Health Research, paper pr05. www.lexjansen.com/pharmasug/2006/publichealthresearch/pr05.pdf

Focus on Energy (2017). "Calendar Year 2016 Evaluation Report – Appendix I: Net Savings Analysis Methodologies." Prepared for the Public Service Commission of Wisconsin by Cadmus Group.
https://focusonenergy.com/sites/default/files/Wisconsin%20Focus%20on%20Energy%20CY%202016%20Appendices.pdf

Fowlie, M.; Wolfram, C. (September 2009). *Evaluating the Federal Weatherization Assistance Program Using a Randomized Encouragement Design (RED).* Presented to Environmental Energy Technologies Division (EETD), Lawrence Berkeley National Laboratory https://eta.lbl.gov/news/events/2009/09/11/estimating-the-impacts-of-low-income-weatherization-assistance-using-a-random

Fowlie, M.; Greenstone, M.; Wolfram, C. (June 2015). "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program."
http://e2e.haas.berkeley.edu/pdf/workingpapers/WP020.pdf#page=1

GDS Associates, Inc.; Nexant; Mondre Energy (2010). *Net Savings: An Overview.* RFP 2009. Prepared for the Statewide Evaluator.

Goldberg, M.; Kademan, E. (1995). Is It Net or Not? A Simulation Study of Two Methods. In Energy Program Evaluation: Uses, Methods, and Results, pp. 459–465. Chicago, IL: National Energy Program Evaluation Conference.

Greene, W. (2011). *Econometric Analysis*, 7th Ed., Prentice Hall.

Guo, S.; Fraser, M. (2010). *Propensity Score Analysis: Statistical Methods and Applications*, *SAGE Publications, Inc.* (Note: Chapter 4 provides an updated discussion of the Heckman models to self-selection along with appropriate caveats. This discussion can be found at: www.sagepub.com/upm-data/30234_Chapter4.pdf

Gwet, K.L. (2012). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters*. Gaithersburg, MD: Advanced Analytics, LLC.

Gwet, K.L. (2010). *Inter-Rater Reliability Using SAS: A Practical Guide for Nominal, Ordinal, and Interval Data*. Gaithersburg, MD: Advanced Analytics, LLC.

Haeri, H.; Khawaja, M.S. (2012). *The Trouble with Freeriders*. Public Utilities Fortnightly. www.cadmusgroup.com/wp-content/uploads/2012/11/Haeri-Khawaja-PUF-TroublewithFreeriders.pdf.

Hall, N; Ladd, D.; Khawaja, M.S. (2013). "Setting Net Energy Impact Baselines: Building Reliable Evaluation Approaches." Paper presented at the 2013 International Energy Program Evaluation Conference, Chicago, IL.

Heckman, J.J. (1979). "Sample Selection Bias as a Specification Error" *Econometrica* 47(1): 153–161.

Ho, D.; Imai, K.; King, G; Stuart, E. (2007). "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Policy Analysis* 15(3); pp.199–236.

Hoefgen, L.; Clendenning, G.; Osman, A.; Keating, K.; Vine, E.; Lee, A.; Stewart, J.; Stoops, J. (2011). "Finding and Counting Market Effects: A New Construction Program Example." Paper presented at the 2011 International Energy Program Evaluation Conference, Chicago, IL.

Hogan, K. (2015). "Getting it Right: Weatherization and Energy Efficiency are Good Investments." Office of Energy Efficiency and Renewable Energy. https://energy.gov/eere/articles/getting-it-right-weatherization-and-energy-efficiency-are-good-investments

Illinois Energy Efficiency Stakeholder Advisory Group (2016). *Illinois Statewide Technical Reference Manual for Energy Efficiency, Version 5.0*. Volume 4: Cross-Cutting Measures and Attachments. February 11, 2016. http://www.ilsag.info/il_trm_version_5.html

Imbens, G.; Lemieux, T. (2010). "Regression Discontinuity Designs: A Guide to Practice." *Journal of Economic Literature* (48); pp. 281–355.

Indianapolis Power and Light (2015). *Demand-Side Management Evaluation Final Report Appendix A: Self-Report Net-to-Gross Evaluation Methodology*. Prepared by the Cadmus Group and Opinion Dynamics. http://www.oracle.com/us/industries/utilities/cadmus-indianapolis-power-light-3697534.pdf

Itron, Inc. (2010). 2006-2008 Evaluation Report for PG&E Fabrication, Process and Manufacturing Contract Group. Prepared for the California Public Utilities Commission, Energy Division. www.calmac.org/publications/PG%26E_Fab_06-08_Eval_Final_Report.pdf; www.calmac.org/publications/PG%26E_Fab_06-08_Eval_Final_Report_Appendices.pdf

Keating, K. (2009). “Freeridership Borscht: Don’t Salt the Soup.” Paper presented at the 2009 International Energy Program Evaluation Conference.

KEMA, Inc. (2010). *Final Evaluation Report: Upstream Lighting Program*, Volume 1. CALMAC Study ID: CPU0015.01; Prepared for: California Public Utilities Commission, Energy Division. Prepared by: KEMA, Inc., Prime Contractor: The Cadmus Group, Inc. www.calmac.org/publications/FinalUpstreamLightingEvaluationReport_Vol1_CALMAC_3.pdf

Keneipp, M., Meurice, J.; Alspector, D.; Sutter, M.; Krause, R.; Hines, T. (2011). “Getting MIF’ed: Accounting for Market Effects in Residential New Construction Programs.” Paper presented at the 2011 International Energy Program Evaluation Conference. Boston, MA.

Kennedy, P. (2008). *A Guide to Econometrics*, 6th Edition. Wiley-Blackwell, April.

Khawaja, M.S.; Rushton, J.; Keeling, J. (2013). “Sample Design Cross-Cutting Protocols.” Chapter 11 in *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*, www.nrel.gov/docs/fy13osti/53827.pdf.

Loughran David S.; Kulick, J. “Demand-Side Management and Energy Efficiency in the United States”. The Energy Journal 25 (1), pp. 19-41. 2004.

Mahalanobis, P. (1936). “On the Generalised Distance in Statistics.” *Proceedings of the National Institute of Sciences of India 2* (1):49–55. refhub.elsevier.com/S0021-9290(13)00445-4/sbref22

McMenamin, J.S. (2008). *Defining Normal Weather for Energy and Peak Normalization.* Itron, Inc. https://www.itron.com/PublishedContent/Defining%20Normal%20Weather%20for%20Energy%20and%20Peak%20Normalization.pdf.

McKenzie, D. (2009). “Impact Assessments in Finance and Private Sector Development -- What Have We Learned and What Should We Learn?” The World Bank Development Research Group, Policy Research Working Paper 4944, May. See: https://openknowledge.worldbank.org/bitstream/handle/10986/4137/WPS4944.pdf?sequence=1

Megdal, L.; Patil, Y.; Gregoire, C.; Meissner, J.; Parlin, K. (2009). “Feasting at the Ultimate Enhanced Freeridership Salad Bar.” Paper presented at the International Energy Program Evaluation Conference, Portland, OR. www.anevaluation.com/pubs/Salad%20Bar%202009%20IEPEC%20paper%205-12-09.pdf.

Messenger, M.; Bharvirkar, R.; Golemboski,B.; Goldman, C.; Schiller, S. (2010). *Review of Evaluation, Measurement and Verification Approaches Used to Estimate the Load Impacts and Effectiveness of Energy Efficiency Programs*. Lawrence Berkeley National Laboratory. http://emp.lbl.gov/sites/all/files/lbnl-3277e.pdf .

Miller, K. (2011). “Cognitive Interviewing.” In *Question Evaluation Methods: Contributing to the Science of Data Quality,* pp. 51–76. Jennifer Madans, Kristen Miller, Aaron Maitland, and Gordon Willis (Eds.) Hoboken, NJ: John Wiley & Sons, Inc.

Mort, D. (2013). "Metering Cross-Cutting Protocols." Chapter 9 of *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*." www.nrel.gov/docs/fy13osti/53827.pdf.

Navigant (2013a). *Custom Free Ridership and Participant Spillover Jurisdictional Review*. Prepared for the Sub-Committee of the Ontario Technical Evaluation Committee, May. www.ontarioenergyboard.ca/documents/TEC/Evaluation%20Studies%20and%20Other%20Reports/Ontario%20NTG%20Jurisdictional%20Review%20-%20Final%20Report.pdf.

Navigant (2013b). *Impact and Process Evaluation of the U.S. Department of Energy's Wind Powering America Initiative.* Prepared for: Department of Energy Office of Energy Efficiency and Renewable Energy, Final Report, DOE/EE-0897, May. http://www1.eere.energy.gov/analysis/pdfs/wind_powering_america_evaluation_2013.pdf

Navigant (2016). *ComEd/Citizens Utility Board Energy Saver Program Evaluation Report*. December. http://ilsagfiles.org/SAG_files/Evaluation_Documents/ComEd/ComEd_EPY8_Evaluation_Reports_Final/ComEd_PY8_CUB_Energy_Saver_Evaluation_Report_2016-12-21_final.pdf

NEEP (2012). "*Regional Net Savings Research, Phase 2: Definitions and Treatment of Net and Gross Savings in Energy and Environmental Policy.*" Submitted to the Northeast Energy Efficiency Partnerships: Evaluation, Measurement, and Verification Forum, by NMR Group, Inc. and Research Into Action, December. http://www.neep.org/net-savings-research-phase-2-0
NEEP (2014). *Decision-Framework for Determining Net Savings Approach – Supplemental Document #2 to Principles and Guidance.* Submitted to the Northeast Energy Efficiency Partnerships: Evaluation, Measurement, and Verification Forum, by Johnson Consulting Group, September. http://www.neep.org/sites/default/files/FINAL%20Supplemental%20Document%202%20Decision%20Framework%20for%20Net%20Savings%20Methods%20Johnon%20Issued%20May%202016.pdf

NEEP (2016). *Gross Savings and Net Savings: Principles and Guidance.* Submitted to the Northeast Energy Efficiency Partnerships: Evaluation, Measurement, and Verification Forum, by Dan Violette (Navigant) and Pam Rathbun (Tetra Tech), April. http://www.neep.org/gross-and-net-savings-principles-and-guidance

New York Department of Public Service (2013a). *Guidelines for Calculating the Relative Precision of Program Net Savings Estimates*. Appendix I.

New York Department of Public Service (2013b). *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approach*. Appendix H.

New York Department of Public Service (2012). *Evaluation Plan Guidance for EEPS Program Administrators,* Update #3, Appendix F. Albany, New York.

Nexant (2017). 2016 Load Impact Evaluation of California Statewide Nonresidential Critical Peak Pricing Program and SCTD Commercial Thermostats. Prepared for Joint CA IOUs, *CALMAC ID SDG0307, April.* http://www.calmac.org/AllPubs.asp

NMR Group, Inc. (2017). *MA RLPNC 16-6: Lighting Shelf Stocking.* Report submitted to the Massachusetts Energy Efficiency Program Administrators. http://ma-eeac.org/wordpress/wp-content/uploads/RLPNC-16-6-%E2%80%93-Lighting-Shelf-Stocking-Report.pdf

NMR Group, Inc. and Research Into Action, Inc. (2010). *Net Savings Scoping Paper.* Revised Draft prepared for Northeast Energy Efficiency Partnerships: Evaluation, Measurement, and Verification Forum. http://www.neep.org/net-savings-scoping-paper-1

NMR Group, Inc. and Tetra Tech (2011). *Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches*. Final report prepared for the Massachusetts Program Administrators. http://ma-eeac.org/wordpress/wp-content/uploads/Cross-Cutting-Net-to-Gross-Methodology-Study-for-Residential-Programs-Suggested-Approaches-Final-Report.pdf

NMR Group, Inc.; KEMA; Cadmus Group, Inc.; Tetra Tech (2011). *Massachusetts ENERGY STAR Lighting Program: 2010 Annual Report.* Final Report prepared for Energy Efficiency Advisory Council Consultants, Cape Light Compact, NSTAR, National Grid, Unitil, and Western Massachusetts Electric. https://www.efis.psc.mo.gov/mpsc/commoncomponents/viewdocument.asp?DocId=935690223

Nonresidential Net-To-Gross Ratio Working Group (2012). *Methodological Framework for Using the Self-Report Approach to Estimating Net-to-Gross Ratios for Nonresidential Customers.* Prepared for the Energy Division, California Public Utilities Commission.

Oak Ridge National Laboratory (1991). *Handbook to DSM Program Evaluation.* Eric Hirst and John Reed, eds., NTIS Pubs., Washington, DC, # ORNL/CON -336, December.

Optimal Energy, Wirtshafter Associates, Inc. (2015). Delaware Evaluation Framework, Prepared for the Department of Natural Resources and Environmental Control. http://www.dnrec.delaware.gov/energy/information/otherinfo/Documents/EEAC/DRAFT%20Delaware%20Evaluation%20Framework.pdf

Parfomak, P.; Lave, L. (1996). "How Many Kilowatts Are in a Negawatt? Verifying the Ex-Post Estimates of Utility Conservation Impacts at a Regional Level." *Energy Journal* 17 (4).

Peters, J.; McRae, M. (2008). "Free-Ridership Measurement Is Out of Sync with Program Logic…or, We've Got the Structure Built, but What's Its Foundation." In Proceedings of the 2008 ACEEE Summer Study on Energy Efficiency in Buildings, Washington, DC. www.aceee.org/files/proceedings/2008/data/papers/5_491.pdf.

PG&E (2013). Evaluation of Pacific Gas and Electric Company's Home Energy Report Initiative for the 2010-2012 Program. Prepared by Freeman, Sullivan Co. Available on the CALMAC.org website: www.calmac.org/publications/2012_PGE_OPOWER_Home_Energy_Reports__4-25-2013_CALMAC_ID_PGE0329.01.pdf.

PG&E (October 2016). Industry Standard Practice (ISP) Study of Portable Irrigation Piping Systems. Prepared by BASE Energy, Inc. Prepared for Dr. Tengfang Xu, Industrial, Agricultural and Water Program, Pacific Gas & Electric Company.

PowerStream (July 2016). Advantage Power Pricing Pilot -- Impact and Process Evaluation. Prepared by Navigant. Reference No.: 172240. https://www.powerstream.ca/attachments/PowerStream_APP_Summer_2015_Evaluation.pdf

Prahl, R.; Ridge, R.; Hall, N.; Saxonis, W. (2013). "The Estimation of Spillover: EM&V's Orphan Gets a Home." In Proceedings of the 2013 International Energy Program Evaluation Conference, August.

Provencher, B; Vittetoe-Glinsmann, B.; Dougherty, A.; Randazzo, K.; Moffitt, P., Prahl, R. (2013). *Some Insights on Matching Methods in Estimating Energy Savings for an Opt-In, Behavioral-Based Energy Efficiency Program.* 2013 International Energy Program Evaluation Conference, Chicago.

Provencher, B.; Glinsmann, B. (February 2013). *Evaluation Report: Home Energy Reports – Plan Year 4.* Prepared for Commonwealth Edison Company.

Puget Sound Energy (2012). *Home Energy Reports Program: Three Year Impact, Behavioral and Process Evaluation*. Prepared for: Puget Sound Energy, Prepared by: KEMA, Inc. https://conduitnw.org/_layouts/Conduit/FileHandler.ashx?RID=849

Ridge, R. (1997). *Errors in Variables: A Close Encounter of the Third Kind*. In Proceedings of the 1997 International Energy Program Evaluation Conference. August, Chicago, IL.

Ridge, R.; Baker, M.; Hall, N.; Prahl, R.; Saxonis, W. (2013). "Gross Is Gross and Net Is Net: Simple, Right?" Paper presented at the 2013 International Energy Program Evaluation Conference, Chicago, IL.

Ridge, R.; Willems, P.; Fagan, J.; Randazzo, K. (2009). "The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating Net-to-Gross Ratio." Paper presented at the 2009 Energy Program Evaluation Conference, Portland, OR.

Rosenberg, M.; Hoefgen, L. (2009). *Market Effects and Market Transformation: Their Role in Energy Efficiency Program Design and Evaluation.* Prepared for the California Institute for Energy and the Environment and the California Public Utilities Commission Energy Division.

RTF (2012). *Guidelines for the Development and Maintenance of RTF Savings Estimation Methods.* NW Council, Released December 4. http://rtf.nwcouncil.org/subcommittees/deemed/

RTF (December 2015). Roadmap for the Assessment of Energy Efficiency Measures. Meeting Website link: https://rtf.nwcouncil.org/meeting/rtf-meeting-december-8-2015 Roadmap document link: https://nwcouncil.app.box.com/v/OperativeGuidelines-20151208

RTF (April 2017). Guidelines Updates Presentation. By Ms. Jennifer Light. https://nwcouncil.app.box.com/v/GuidelinesAprilRTFPres

Ruegg, R.; Jordan, G. (2007). *Overview of Evaluation Methods for R&D Programs: A Directory of Evaluation Methods Relevant to Technology Development Programs*. Prepared for the U.S. Department of Energy: Office of Energy Efficiency and Renewable Energy. https://www1.eere.energy.gov/analysis/pdfs/evaluation_methods_r_and_d.pdf

Rufo, M. (2009). "Evaluation and Performance Incentives: Seeking Paths to (Relatively) Peaceful Coexistence." In Proceedings of the International Energy Program Evaluation Conference, Portland, OR, August.

Sacramento Municipal Utilities District (2011). *Evaluation Report: OPOWER Sacramento Municipal Utilities District (SMUD) Pilot Year2*. Prepared by Navigant. February. http://www.oracle.com/us/industries/utilities/opower-smud-pilot-year-2-3631985.pdf

Sacramento Municipal Utilities District (2013). Load Impact Results from SMUD's Smart Pricing Options Pilot. Prepared by Freeman Sullivan & Co. for Sacramento Municipal Utility District – SMUD contact Ms. Lupe Jimenez

Scriven, M. (2009). "Demythologizing Causation and Evidence." In *What Counts as Credible Evidence in Applied Research and Evaluation Practice*. Stewart I. Donaldson, Christina A. Christie, and Melvin Mark (Eds.). Los Angeles, CA: SAGE Publications.

Sebold, F.D.; Fields, A.; Skumatz, L.; Feldman, S.; Goldberg, M.; Keating, K.; Peters, J. (2001). *A Framework for Planning and Assessing Publicly Funded Energy Efficiency*. www.calmac.org/events/20010301PGE0023ME.pdf

SEE Action (2012a). *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 Berkley National Laboratory. https://www4.eere.energy.gov/seeaction/system/files/documents/emv_behaviorbased_eeprograms.pdf

SEE Action (2012b). *Energy Efficiency Program Impact Evaluation Guide*. Prepared by Steven R. Schiller, Schiller Consulting, Inc. See: https://www4.eere.energy.gov/seeaction/system/files/documents/emv_ee_program_impact_guide_0.pdf

Shadish, W.R., Cook, T.D.; Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin Company.

Southern California Edison (2014). *Edison SmartConnect Demand Response and Energy Conservation Annual Report*,. http://www3.sce.com/sscc/law/dis/dbattach5e.nsf/0/98EF18FDB79332B688257CCA007BFA88/$FILE/R1309011%20DR%20OIR%20-%20SCE%20Compliance%20Filing%20Pursuant%20to%20Decision%2008-09-039.pdf

Stryker, A.; Gaffney, K. (2013). "Why the Light Bulb Is No Longer a Textbook Example for Price Elasticity: Results from Choice Experiments and Demand Modeling Research." In Proceedings of the International Energy Program Evaluation Conference, Chicago, IL, August.

Stuart, E.A. (2010). "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25(1):1–21.

TecMarket Works; The Cadmus Group; Opinion Dynamics Corporation; Integral Analytics; Building Metrics; Energy Efficient Homes Midwest (2012). *Indiana Evaluation Framework*. Prepared for the Indiana Demand Side Management Coordination Committee.

Tetra Tech; NMR Group; DNV GL (2017). *Net-to-Gross Methodology Research (TXC08)*. Massachusetts Program Administrators. http://ma-eeac.org/wordpress/wp-content/uploads/Net-to-Gross-Methodology-Research.pdf

Tetra Tech; NMR Group; and DNV GL (2017). *Top-Down Modeling Extended Methods Review*. Massachusetts Program Administrators.

Tetra Tech.; KEMA; NMR Group, Inc. (2011). *Cross-Cutting (C&I) Free Ridership and Spillover Methodology Study Final Report.* Massachusetts Program Administrators. http://ma-eeac.org/wordpress/wp-content/uploads/Massachusetts-PAs-Cross-Cutting-CI-Free-ridership-and-Spillover-Methodology-Study.pdf

Titus, E.; Michals, J. (2008). "Debating Net Versus Gross Impacts in the Northeast: Policy and Program Perspectives." ACEEE Summer Study on Energy Efficiency in Buildings, (5): 312–323. https://www.aceee.org/files/proceedings/2008/data/papers/5_429.pdf

U.S. Department of Energy (2010). *Guidance Document #7: Topic: Design and Implementation of Program Evaluations that utilize Randomized Experimental Approaches.* Smart Grid Investment Grant Technical Advisory Group, November. https://www.smartgrid.gov/files/cbs_guidance_doc_7_randomized_experimental_approaches.pdf

Violette, D. (2013). "Persistence and Other Evaluation Issues Cross-Cutting Protocols." Chapter 13 in *Uniform Methods Project for Determining Energy Efficiency Program Savings for Specific Measures*. NREL/SR-7A30-53827, April. See: www1.eere.energy.gov/wip/pdfs/53827-13.pdf

Violette, D.; Barkett, B.; Schare, S.; Skumatz, L.; Dimetrosky, S. (2005). *Commercial/Industrial Performance Program (CIPP) Market Characterization, Market Assessment And Causality Evaluation.* Prepared for NYSERDA, Jennifer Ellefsen, Project Number 7721, March.

Violette, D.; Brakken, R.; Schon, A.; Greer, J. (1993). *Statistically-Adjusted Engineering Estimates: What Can the Evaluation Analyst Do About the Engineering Side Of the Analysis?* Published in the *Proceedings of the 1993 International Energy Program Evaluation Conference (IEPEC)*.

Violette, D.; Keneipp, M.; Ozog, M., Stern, F. (1991). *Impact Evaluation of Demand-Side Management Programs — Volume 1: A Guide to Current Practice*. Electric Power Research Institute Pubs., Palo Alto, CA, #EPRI CU-7179, February.

Violette, D.; Ozog M.; Cooney, K. (2003), Retrospective Assessment of the Northwest Energy Efficiency Alliance -- Findings and Report. *Prepared for:* Northwest Energy Efficiency Alliance, Ad Hoc Retrospective Committee, December 8. See: http://www.theboc.info/pdf/Eval-BOC_SummittBlue_NEEA_2003.pdf

Violette, D.; Provencher, B. (2012). *Review of a Top Down Evaluation Study: Rivers & Jaccard (2011).* Prepared for BC Hydro, Navigant Consulting, Inc., April.

Violette, D.; Rogers, B. (2012). *A Sampling Methodology for Custom C&I Programs*. Prepared by Navigant Consulting, Inc., prepared for the Ontario Natural Gas Technical Evaluation Committee, Ontario Energy Board, November. Report available at Ontario Energy Board website: https://www.researchgate.net/publication/279531668_MAIN_REPORT_-_A_Sampling_Methodology_for_Custom_CI_Programs_Technical_Evaluation_Committee

Violette, D.M.; Provencher, B.; Sulyma, I. (2012). "Assessing Bottom-Up and Top-Down Approaches for Assessing DSM Programs and Efforts." In International Energy Program Evaluation Conference Proceedings, Rome, June.

Weiss, C. (1997). "Theory-Based Evaluation: Past, Present, and Future" Special Issue: Progress and Future Directions in Evaluation: Perspectives on Theory, Practice, and Methods, New Directions in Evaluation, Volume 1997, Issue 76.

West, S. (2008). "Alternatives to the Randomized Controlled Trial." *American Journal of Public Health* 98(8). http://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2007.124446

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA, MIT Press Ltd.

Xcel Energy (2016). 2017-19 RE-TOU & RD-TDR Study and Evaluation Plan. November 15 Filing. http://www.dora.state.co.us/pls/efi/EFI.Show_Filing?p_session_id=&p_fil=G_726523

7 Appendix A: Price Elasticity Studies as a Component of Upstream Lighting Net Savings

Studies of upstream changes in the price for residential lighting products have received attention as a way to complement surveys with market actors, or even replace these surveys with econometric models. The way in which price can be viewed as a driver of program savings and the importance of other program components is discussed in Stryker and Gaffney (2013).

Price elasticity studies are currently being applied in several jurisdictions. To date, these studies have focused on residential lighting products and, within that category, mostly on CFL sales. For example, Cadmus (2012b, 2013) and KEMA (2010) tested several different methods for estimating the increase in CFL sales resulting from a program-induced price reduction caused by program activities (markdowns negotiated with retailers and coupons). These two approaches are outlined below.[95]

Cadmus (2012b) examined Efficiency Maine's residential lighting program and Cadmus (2013) examined Wisconsin's Focus on Energy residential lighting program. Both studies used a price elasticity approach. These two studies estimated expected bulb purchases (and associated savings) at prices offered under the program and then the purchases that would have occurred at original retail prices. The difference between these two values was viewed as net savings in this study.

Cadmus (2012b, 2013) used a single equation regression model where the quantity of CFLs purchased was a function of the price of CFLs and a select set of other independent variables. The data used to estimate this equation included package and bulb sales for each retailer, by model number and by week. The dataset does not include information about the consumers who purchased the CFLs, but does contain information about quantities of CFLs sold and retailer prices. Consumer variables desirable in a demand equation would include income and education, but often these variables are not available in the retailers' sales tracking systems.

A regression was estimated relating quantities of CFLs sold by retailer to the price of CFLs that week for each retailer. Other factors such as promotional events were considered in determining consumer purchases. Programmatic factors such as labeling and information dissemination are pervasive throughout the lighting programs and, while potentially important, could not be addressed due to lack of variation across consumer purchases.

These two studies showed an increase in the sales of CFLs as prices decreased due to markdowns negotiated with retailers and discount coupons provided to consumers. The second step of the approach involved estimating what the sales would have been at the higher prices that would have prevailed without the program (that is, the counterfactual scenario).

[95] Both Cadmus and KEMA (now DNV GL) have completed more recent studies using price elasticity approaches for upstream lighting programs. Each incorporates several new features, but the constructs are similar to those discussed in this section. Updated citations for more recent applications are Focus on Energy (2017) and DNV GL (2017).

Considerable effort was made in these price elasticity studies to control for factors other than price that might also affect CFL sales, but it is difficult to show that any method is free of bias. In the case of the Efficiency Maine lighting program, there were three components to the program. Two were linked to price (markdowns and coupons) and a third was linked to overall participation in the Appliance Rebate Program, "with Appliance Rebate Program participants electing to receive a free six-pack of CFL bulbs, via a check-off on the Appliance Rebate Program application form." The third part of the program would have provided CFLs at essentially no cost and it is not clear how this would have factored into the analysis.

Cadmus (2012b, 2013) present several general caveats to the demand equation approach used in the study. First, it acknowledged that "this estimation method has rarely been used in upstream lighting program evaluations as such data generally have been unavailable. As Efficiency Maine … tracked these data and shared them for this evaluation, Cadmus found such econometric demand estimation provided the best method for estimating the program's free-ridership." Second, Cadmus (2013) indicates that it "will continue to look for alternative methods to calculate net-to-gross," and that "the model used for the … 2012 evaluation does not account for spillover."[96]

KEMA (2010) used price variables to estimate net savings in an upstream lighting study. This study had the benefit of a sizeable data collection effort that included consumer surveys. As part of the in-store consumer intercept research, brief interviews were conducted with shoppers who had just made a lighting purchase (revealed preference) as well as "stated preference" surveys with other consumers recruited randomly. Intercept surveys were conducted with 1,463 customers across 378 stores.

KEMA (2010) used three primary types of methods for estimating net savings:

- Supplier and consumer self-report methods
- Econometric models
- Total sales (market-based) approach.

Among the econometric modeling efforts, four econometric models were used:

- Pricing (price formation model)
- Conjoint elasticity
- Revealed preference purchase
- Stated preference purchaser elasticity.

The first two econometric methods—price formation and the conjoint elasticity model—were both needed to produce a net savings estimate. Revealed preference and stated preference models

[96] Cadmus (2012) indicates that spillover is not addressed in this study; however, looking at the overall change in sales in a market caused by price elasticity, has included spillover elements in other studies that use a similar price elasticity approach.

can produce net savings directly. As a result, there were four econometric models, but only three approaches for estimating net savings.

The price formation model estimates the percentage reduction in CFL prices that resulted from program incentives. This is combined with the conjoint analysis, which estimated the corresponding percentage increase in market share/sales that result from a price decrease. This allowed the net savings to be calculated by combining the findings from the pricing study with the conjoint demand elasticity study—in other words, the program induced reduction in prices from the pricing study multiplied by the estimate of change in sales caused by a lower price from the conjoint study.

KEMA (2010) revealed a preference for store intercepts to survey customers that made actual CFL purchases. These customers were asked to indicate how many CFLs they would have bought compared to their actual purchases at double the price they actually paid. Response categories were: (1) the same amount, (2) fewer, and (3) none. Although still based on hypothetical, self-reported responses, the revealed preference respondents may be a more reliable sample because they just made an active purchase decision. However, revealed preference respondents may be somewhat unlikely to indicate they would have paid more for what they just purchased. KEMA (2010) used a random survey of customers, including customers who did not actually purchase a CFL. KEMA (2010) states that the magnitude of the potential bias across these two methods is unknown, "but it is likely that NTG ratio estimates from stated preference respondents are biased downward and NTG ratio estimates from revealed preference respondents are biased upward."

The revealed preference model allowed KEMA to use the store-intercept survey data to model CFL purchase rates with and without program effects. This model was based on a logistic regression to model the probability of buying a CFL rather than an "equivalent" non-CFL as a function of price, displays, customer characteristics, and bulb characteristics, by channel. The fitted models were evaluated under program and non-program conditions. For each channel, the difference between the probability of purchasing CFLs under the program condition and that under the non-program condition was the program-attributable CFL sales share.

In summary, the price elasticity studies completed to date have been limited to residential lighting programs. Cadmus (2012b, 2013) developed a demand model specification based on an examination of alternative specifications. KEMA (2010) developed several approaches for examining the change in CFLs sold as a function of program-induced lower prices. KEMA (2010) concluded that from the econometric approaches, the revealed preference model was the preferred approach. It should be noted that these approaches focus on free-ridership and do not address spillover or longer-term market effects. Currently, several evaluations are using the price-elasticity method to estimate net savings from residential lighting. An expanded literature will likely provide additional confidence in this method for addressing free-ridership from upstream lighting programs, and possibly an expansion of this method to other residential product programs.

Chapter 22: Compressed Air Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in November 2014. The content in this version has been updated.

Nathanael Benton and Patrick Burns
Nexant, Inc.
San Francisco, California

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68577
October 2017

Contract No. DE-AC36-08GO28308

Chapter 22: Compressed Air Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This version supersedes the version originally published in November 2014. The content in this version has been updated.

Nathanael Benton and Patrick Burns
Nexant, Inc.
San Francisco, California

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68577
October 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter authors wish to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Jeff Bartels and Patrick Burns of Nexant
- Jeff Cropp and Joel Zahlan of Cadmus
- Keith Downes of Navigant.

Suggested Citation

Benton, N.; Burns, P. (2017). *Chapter 22: Compressed Air Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68577. http://www.nrel.gov/docs/fy18osti/68577.pdf

Acronyms

CAGI	Compressed Air and Gas Institute
CCAF	Compressor Control Type Adjustment Factor
CFM	cubic feet per minute
ECM	electronically commutated motor
gal	gallons
hp	horsepower
kW	kilowatt
kWh	kilowatt hour
psi	pounds per square inch
psia	pounds per square inch absolute
psig	pounds per square inch gauge
RMS	root mean square
SCFM	standard cubic feet per minute
VSD	variable-speed drive

Protocol Updates

The original version of this protocol was published in November 2014.

This chapter has been updated to incorporate the following revisions:

- Added guidance on establishing baseline assumptions for new construction and replace on failure applications.
- Added performance curves for load/unload controlled rotary screw compressors with 5 gallons per cubic feet per minute (CFM) receiver capacities and demonstrated the appropriate method for developing unique performance curves through interpolation.
- Provided additional measurement and verification guidance on how to develop average hourly estimates of compressed air demand including a discussion on the advantages of incorporating day types into a CFM-bin analysis.
- Outlined scenarios where historical trend data can be used in lieu of independent metering.
- Added discussion on the appropriate use of ultrasonic leak detectors and the importance of pre and post survey leak-down tests in estimating reductions in air loss (CFM).
- Added default performance curves for centrifugal air compressors.

Table of Contents

List of Figures

List of Tables

1 Measure Description

Compressed-air systems are used widely throughout industry for many operations, including pneumatic tools, packaging and automation equipment, conveyors, and other industrial process operations. Compressed-air systems are defined as a group of subsystems composed of air compressors, air treatment equipment, controls, piping, pneumatic tools, pneumatically powered machinery, and process applications using compressed air. A compressed-air system has three primary functional subsystems: supply, distribution, and demand.

Air compressors are the primary energy consumers in a compressed-air system and are the primary focus of this protocol.[1] The two compressed-air energy efficiency measures specifically addressed in this protocol are:

- High-efficiency/variable speed drive (VSD) compressor replacing modulating, load/unload, or constant-speed compressor
- Compressed-air leak survey and repairs.

This protocol provides direction on how to reliably verify savings from these two measures using a consistent approach for each.

1.1 High-Efficiency/Variable-Speed Drive Compressor Replacing Modulating Compressor

This measure pertains to the installation of a rotary screw compressor with a VSD. Most incentive programs and technical reference manuals use a baseline system definition of a standard modulating compressor with blowdown valve. The energy-efficient compressor is typically defined as an oil-flooded, rotary-screw compressor with variable-speed control.

This measure is frequently offered for the replacement of an existing unit at the end of its useful life or for the installation of a new system in a new building (i.e., time of sale).

Several control methods are available for air compressors, and control methods greatly affect the overall operating efficiency of a compressor. To accurately estimate energy savings, it is important to know the baseline method of control. A brief description of each common control method is provided below.

1.1.1 Reciprocating – On/Off Controls

The simplest method of control is to use an on/off control to start and stop a compressor to maintain system pressure. The compressor starts and generates air when the pressure falls below a certain set point, and it turns off when pressure is above a certain set point. Using an on/off control is an efficient way to ensure the compressor is either fully loaded or off; however, this form of control is only suitable for small compressors (typically less than approximately 5

[1] As discussed in the "Considering Resource Constraints" section of the UMP Chapter 1: Introduction, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities. http://www.sba.gov/category/navigation-structure/contracting/contracting-officials/small-business-size-standards

horsepower [hp] and most common in residential settings). This method of control is uncommon in industrial settings.

1.1.2 Reciprocating – Load/Unload Control

Reciprocating compressors can be unloaded by holding open the inlet valve. Air is still pushed in and out of the compression chamber, but it is not compressed and discharged to the system. Depending on the number of cylinders and controls, the system may have multiple loading steps, such as 0%–50%–100% or 0%–25%–50%–75%–100%. Some compressors have a variable clearance volume, which impacts the amount of compressed air discharged at the end of the piston stroke and allows for additional capacity adjustment. Regardless of the specific type or steps of control, the standard performance curve shown in Table 2, later in this document, represents the energy usage.

1.1.3 Rotary-Screw – Inlet Valve Modulation/Inlet Throttling

Inlet valve modulation throttles off the air inlet to a compressor as discharge pressure rises above the set point pressure. The part-load performance of modulating compressors is relatively poor. Some modulation-controlled machines may be adjusted to fully unload if capacity reduces to a certain level, such as 40%. This reduces energy consumption compared to modulation-only compressors but requires the use of air storage receivers to meet demand when in the fully unloaded state.

1.1.4 Rotary-Screw – Load/Unload Control

Load/unload controls require significant storage receiver volume and operate a compressor at full capacity until the unload pressure (cutout) set point is reached. The compressor then unloads and blows down the oil separator and operates at minimum power while producing no air. Oil-free screw compressors nearly instantly unload due to no oil separator blowdown. The air loss associated with blowing down the oil separator is also eliminated.

1.1.5 Rotary-Screw – Variable-Displacement Control

Variable-displacement controls change compressor capacity by opening ports in the compressor that limit the amount of the cylinder or air-end that is used for compression. This can be implemented in either discrete steps (e.g., 50%, 75%, and 100%) or by continuously varying capacity. Compressor-specific power is typically good within the variable displacement range, but these compressors typically have a limited turndown range. At minimum turndown, the compressor commonly uses inlet modulation to further reduce flow, resulting in poor specific power, or kilowatt (kW) per CFM.

1.1.6 Rotary-Screw – Variable-Speed Control

VSD or variable-frequency drive compressor controls use an integrated variable frequency alternating current or switched-reluctance direct current drive to control the electrical signal to the motor and, in turn, vary the speed of the motor and compressor. Compressors equipped with VSD controls continuously adjust the drive motor speed to match variable demand requirements. VSD compressors typically have an excellent turndown range and efficiently produce air over the entire range of operating speeds. Below the minimum turndown speed, the compressor typically cycles between off and minimum-load states. This method of control is typically the high-efficiency case and not the base case.

1.1.7 Centrifugal Controls

Most centrifugal compressors use a form of inlet throttling to vary capacity. Inlet butterfly valve and inlet guide vanes are both similar methods of control that reduce flow while also reducing power. Due to limitations in centrifugal compressor design, flow can only be reduced to a minimum level before surging occurs. To meet system flow below the throttling range, which is typically below approximately 70% of full load capacity, variable demands are met by the compressor by operating at the minimum throttled position and blowing off excess air produced through a blow-off valve. Therefore, most centrifugal compressors use a constant amount of power below the throttle limit regardless of actual demand. The standard curves shown in Table 2 are reflective of these common methods of control.

Another method of control used for centrifugal compressors is inlet throttling with unloading. Some centrifugal compressors can unload and recycle compressed air back to the compressor inlet instead of blowing off, and wasting, generated air. This control method can be more efficient, but loading cycles do not allow for constant system pressure control.

A newer centrifugal compressor type uses a high-speed variable-speed rotor supported by magnetic bearings. The compressor varies speed to meet loads within the throttling range and unloads to a reduced speed instead of blowing off excess air. This type of control can be highly efficient, although it is not a compressor type commonly available at the time of this writing. It is important to note that variable-frequency drives cannot be retrofitted to existing fixed-speed centrifugal compressors; a special type of compressor is needed to utilize this advanced method of control.

For all centrifugal compressors, obtaining the actual performance curve is recommended as the performance of different compressor models varies significantly.

1.2 Compressed-Air Leak Survey and Repairs

Leaks are a significant cause of wasted energy in a compressed-air system and can develop in many parts of a compressed air system. The most common problem areas are couplings; hoses; tubes; fittings pipe joints quick disconnects; filters, regulators, and lubricators; condensate traps; valves; flanges; packings; thread sealants; and other point-of-use devices.

Leakage rates are a function of the supply pressure, typically quantified in standard cubic feet per minute (SCFM), and proportional to the square of the orifice diameter (hole or crack size).

There are three common methods of compressed-air leak detection: auditory and sensatory observation, soapy water test, and ultrasonic leak detection. The industry standard and best practice is ultrasonic leak detection. This relies on the ability of specialized directional microphones and amplifiers to detect high-frequency noise generated by the turbulent flow of compressed air escaping a compressed-air system through an orifice or crack. The high-frequency sound produced by a compressed-air leak is both directional and localized to the source.

2 Application Conditions of Protocol

2.1 High-Efficiency/Variable-Speed Drive Compressor Replacement Measures

Demand-side management programs typically offer a prescriptive compressor replacement measure. Many programs and technical reference manuals assume the baseline compressor system to be a modulating, load/unload, or constant-speed compressor. New energy-efficient compressors are assumed to be VSD controlled.

Incentives for air compressor replacements are typically paid on a dollar-per-compressor-horsepower basis, dollar-per-kilowatt hour-saved basis, or a fixed percentage of project cost. Common eligibility requirements for compressor replacement measures include:

- The air compressor must be a primary system component and not a backup system component.
- Replaced equipment must be removed or the customer must attest that the baseline system, if remained connected, will be used only for emergency backup purposes and will rarely (if ever) operate.
- Only one VSD compressor per system is eligible for incentive.

This measure is commonly offered for retrofit (or early replacement) projects and new construction or replace on burnout/time-of-sale projects. For a new construction project or if the baseline unit has failed or is near the end of its useful life, the baseline efficiency should be determined from:

- The market industry standard/common practice for the given baseline control type
- Compressed Air and Gas Institute (CAGI) performance sheet data for an equivalently sized new compressor with load/unload or modulating controls.

This protocol is also applicable to projects involving the addition of a VSD controlled trim compressor to a multiple compressor central plant and to projects where an existing air compressor is retrofitted with an add-on VFD.

2.2 Compressed-Air Leak Surveys and Repairs

Compressed-air leak surveys are typically performed by a program-approved third party or a trade ally. Programs typically establish specific guidelines for conducting the survey and reporting the findings.

Energy savings from compressed-air system repairs are determined by multiplying the estimated reduction in compressed air loss in SCFM by the power input per CFM (also known as efficacy) of the air compressor serving the system for the range of loading experienced by the system.

Incentives are typically paid as the least of:

- A fixed dollar amount per rated compressor horsepower

- Full reimbursement for the cost of the leak survey
- A program-defined maximum, not-to-exceed dollar amount.

3 Savings Calculations

This section describes the calculation methods for estimating gross savings from compressed air projects.

3.1 Savings Calculations for Installing a High-Efficiency Air Compressor

3.1.1 Compressor Power at Full Load

Energy use reduction for all compressor projects can be calculated by the difference between the energy consumed in the baseline operation minus the energy consumed in the post-retrofit operation. Generally, information is required for compressor capacity in both the baseline and post-retrofit scenarios. Appropriate adjustments are made to ensure the flow profile is equivalent between pre- and post-retrofit conditions unless demand improvements have been made that result in a change in the flow profile.

Compressor power at full load can be calculated as follows:

$$\text{Full Load kW}_{rated} = \frac{(\text{Compressor hp}) \times LF_{rated} \times (0.746\ \text{kW/hp})}{(\eta_{motor})} \tag{1}$$

$$\text{Full Load kW}_{rated} = \frac{(\text{Compressor hp}) \times LF_{rated} \times (0.746\ \text{kW/hp})}{(\eta_{motor}) \times (\eta_{VSD})} \tag{2}$$

where:

Compressor hp	= compressor horsepower, nominal rating of the prime mover (motor)
0.746	= horsepower to kW conversion factor
η_{motor}	= motor efficiency (%)
η_{VSD}	= variable-speed drive efficiency (%)
LF_{rated}	= load factor of compressor at full load (typically 1.0 to 1.2)

VSDs have losses, just like other electronic devices that transform voltage. VSD efficiency decreases with decreasing motor load. The decline in efficiency is more pronounced with drives of smaller horsepower ratings. VSD efficiencies typically range from 94% to 97% depending on the load and compressor horsepower (DOE 2012).

Alternatively, full load power may be available from manufacturers or CAGI performance sheet data. Measuring full- and part-load power is even more accurate for a specific site.

Air compressor full load performance values provided on CAGI data sheets are reported at standard atmospheric conditions (14.7 pounds per square inch absolute [psia] at sea level). Typically, air compressor operating conditions will differ from these standard values, so these values must be corrected to actual operating conditions. The full-load kW is influenced by site elevation and the compressor operating pressure.

The following expressions are used to correct the compressor full-load performance based on site-specific conditions.

$$\mathrm{kW_{adjusted}} = \mathrm{Full\ Load\ kW_{rated}} \times \frac{\left[\left(\frac{\mathrm{P_{discharge}} + \mathrm{P_{alt}}}{\mathrm{P_{alt}}}\right)^{\frac{0.395}{1.395}} - 1\right]}{\left[\left(\frac{\mathrm{P_{rated}} + 14.7}{14.7}\right)^{\frac{0.395}{1.395}} - 1\right]} \tag{3}$$

where:

Full Load $\mathrm{kW_{rated}}$	= full-load kW of air compressor at full load capacity and pressure (per CAGI data sheet or manufacturer specifications)
$\mathrm{P_{discharge}}$	= actual system discharge pressure (psig)
$\mathrm{P_{alt}}$	= atmospheric pressure based on site elevation above sea level (psia)
$\mathrm{P_{rated}}$	= pressure at rated flow (psig) per CAGI data sheet or manufacturer specified design inlet pressure
14.7	= standard atmospheric conditions (psia) at sea level
(0.395/1.395)	= based on the ratio of specific heat for air at standard atmospheric conditions and isentropic compression with constant specific heats

A common rule of thumb for systems in the 80 to 140 pounds per square inch gauge (psig) range is: for every 2 pounds per square inch (psi) increase (or decrease) in discharge pressure, energy consumption will increase (or decrease) by approximately 1% at full output flow. This rule of thumb closely approximates Equation 3 within this range. Outside this range, Equation 3 is preferred. Equation 4 demonstrates how the "rule-of-thumb" adjustment is calculated:

$$\mathrm{kW_{adjusted}} = \mathrm{Full\ Load\ kW_{rated}} \times [1 - (((\underline{\mathrm{P_{rated}} - \mathrm{P_{discharge}}})/2) \times 0.01)] \tag{4}$$

3.1.2 Compressor Power at Part Load

The rated full-load power of a compressor represents the energy use of the system when operating at full load. At part-load conditions, compressor power is generally lower with common control types. To determine power at part load, the part-load fraction, calculated as the supplied CFM divided by the rated CFM for a given compressor, is matched to the percentage of power using an appropriate table (see Table 1 and Table 2). The operating power can then be calculated at a given capacity using Equation 5:

$$\mathrm{kW_{operating}} = \mathrm{kW_{adjusted}} \text{ x } \%\ \mathrm{Power} \tag{5}$$

where:

$kW_{adjusted}$	= Adjusted full-load kW based on actual operating conditions or measured data
% Power	= percentage of power input (%), ratio of the load that the compressor is actually drawing relative to the rated full load Note: % power is not a parameter that can be physically measured, although measuring power and then testing the compressor at full-load will provide the variables needed to calculate percentage of power.

Percentage of power is also influenced by equipment type (reciprocating, rotary screw, etc.) and method of control (throttling, on/off, variable speed, etc.). Table 1 presents typical power versus capacity distributions for rotary screw compressors with multiple control methods. Table 2 presents typical percentage of power versus percentage of capacity curves for centrifugal and reciprocating air compressors. The data in Tables 1 and 2 were developed from standard percentage of power versus percentage of capacity performance curves extracted from Scales and McCulloch (2013) and Smith (2012). Figure 1 shows examples of percentage of power versus percentage of capacity curves for lubricated rotary screw air compressors.

Table 1. Average Percentage of Power Versus Percentage of Capacity for Rotary Screw Compressors with Various Control Methods

(Scales and McCulloch 2013)

Percentage of Capacity	Load/Unload – Oil-Free	Load/Unload (1 gal/CFM)	Load/Unload (3 gal/CFM)	Load/Unload (5 gal/CFM)	Load/Unload (10 gal/CFM)	Inlet Valve Modulation (w/o Blowdown)	Inlet Valve Modulation (w/Blowdown)	Variable Displacement	VSD w/Unloading	VSD w/Stopping
0%	27%	27%	27%	27%	27%	71%	26%	25%	12%	0%
10%	34%	32%	34%	36%	35%	74%	40%	34%	20%	12%
20%	42%	63%	54%	44%	42%	76%	54%	44%	28%	24%
30%	49%	74%	64%	53%	52%	79%	62%	52%	36%	33%
40%	56%	81%	73%	65%	60%	82%	82%	61%	45%	41%
50%	64%	87%	79%	70%	68%	86%	86%	63%	53%	53%
60%	71%	92%	85%	77%	76%	88%	88%	69%	60%	60%
70%	78%	95%	90%	85%	83%	92%	92%	77%	71%	71%
80%	85%	98%	94%	90%	89%	94%	94%	85%	80%	80%
90%	93%	100%	98%	97%	96%	97%	97%	91%	89%	89%
100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

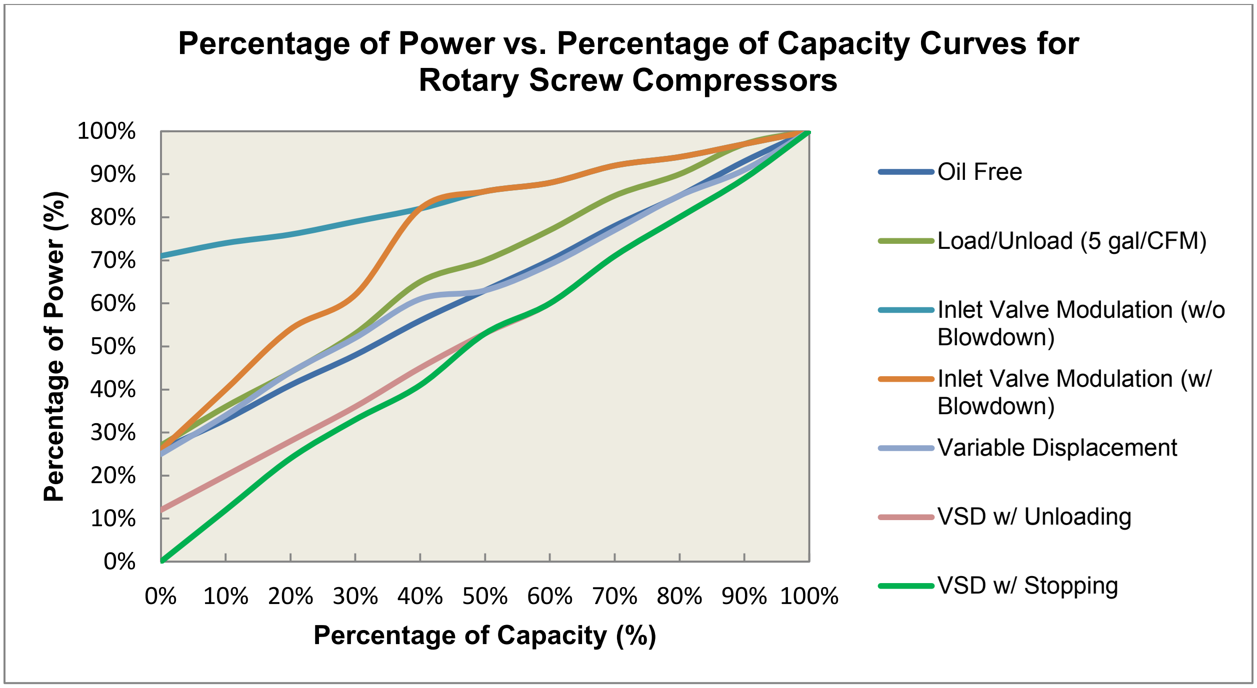


Figure 1. Example Operation Curve (percentage of power versus percentage of capacity curve) for Lubricated Rotary Screw Air Compressor

Table 2. Average Percentage of Power Versus Percentage of Capacity for Reciprocating and Centrifugal Compressors with Various Control Methods

(Compressed Air Challenge)

	Percentage of Power			
Percentage of Capacity	**Reciprocating On/Off Control**	**Reciprocating Load/Unload**	**Centrifugal IBV[a] w/ Blowdown**	**Centrifugal IGV[b] w/ Blowdown**
0%	0%	26%	80%	74%
10%	10%	33%	80%	74%
20%	20%	41%	80%	74%
30%	30%	48%	80%	74%
40%	40%	56%	80%	74%
50%	50%	63%	80%	74%
60%	60%	70%	80%	74%
70%	70%	78%	80%	74%
80%	80%	85%	87%	83%
90%	90%	93%	93%	91%
100%	100%	100%	100%	100%

[a] IBV – Inlet Butterfly Valve Modulation
[b] IGV – Inlet Guide Vane Modulation

In situations where the receiver storage capacity per CFM of supplied air for a load/unload controlled compressed air system does not match one of the default performance curves provided in Tables 1 and 2, it is recommended that a unique profile is developed using the process of interpolation. An example interpolation calculation is provided below.

Example 1: Using Linear Interpolation to Develop Project-Specific Performance Curve for Load/Unload Compressor w/ Compressed Air Storage

Assume the base-case system on a VSD compressor replacement project consists of a load/unload-controlled rotary screw air compressor with a rated flow of 360 SCFM and approximately 1,000 gallons of receiver storage. The ratio of compressed air receiver capacity (gallons [gal]) to supplied SCFM is approximately 2 gal per SCFM. Using interpolation and the values from Table 1 for 1 gal/SCFM and 3 gal/SCFM load/unload systems; approximate the %Power of a 2 gal/SCFM load/unload-controlled system when operating at 60% capacity.

General Formula for Linear Interpolation

$$\%\mathrm{Pwr}_{\mathrm{z,cap\%}} = \%\mathrm{Pwr}_{\mathrm{x,cap\%}} + \left[\frac{\mathrm{gal/CFM_z} - \mathrm{gal/CFM_x}}{\mathrm{gal/CFM_y} - \mathrm{gal/CFM_x}} \times \left(\%\mathrm{Pwr}_{\mathrm{y,cap\%}} - \%\mathrm{Pwr}_{\mathrm{x,cap\%}}\right)\right] \quad (6)$$

where:

%cap	= specified operating point (% capacity)
$\%\mathrm{Pwr}_{z,cap\%}$	= % power of z gal/CFM system at specified % capacity
$\%\mathrm{Pwr}_{x,cap\%}$	= % power of x gal/CFM system at specified % capacity
$\%\mathrm{Pwr}_{y,cap\%}$	= % power of y gal/CFM system at specified % capacity
$\mathrm{gal/CFM}_x$	= lower bound receiver capacity
$\mathrm{gal/CFM}_y$	= upper bound receiver capacity
$\mathrm{gal/CFM}_z$	= receiver capacity of subject system being evaluated

Using the default performance curves for 1 gal/CFM and 3 gal/CFM load/unload compressed air systems and the known receiver capacities, we can approximate the %Power of a 2 gal/CFM system while operating at 60% capacity as follows.

%cap	= 60%	
$\%\mathrm{Pwr}_{x,cap\%}$	= $\%\mathrm{Pwr}_{1\ gal,60\%}$	= 92% (from Table 1)
$\%\mathrm{Pwr}_{y,cap\%}$	= $\%\mathrm{Pwr}_{3\ gal,60\%}$	= 85% (from Table 1)
$\mathrm{gal/CFM}_x$	= 1 gal/CFM	
$\mathrm{gal/CFM}_y$	= 3 gal/CFM	
$\mathrm{gal/CFM}_z$	= 2 gal/CFM	

$$\%\text{Load}_{2\ gal,60\%} = 0.92+\left[\left(\frac{2\text{-}1}{3\text{-}1}\right)\times\left(0.85\text{-}0.92\right)\right]$$

$$\%\text{Load}_{2\ gal,60\%} = 0.845$$

This process can be repeated for all other common operating points (percentage of capacity values) relevant to the given project or a unique performance curve can be developed by interpolating %Power values for the full range of 0% to 100% in 10% increments (as shown in Table 3)."

Table 3. Interpolated Percentage of Power Versus Percentage of Capacity Curve for Rotary Screw Compressor with Load/Unload Controls and Receiver Capacity of 2 gal/CFM

Percentage of Capacity	Percentage of Power Load/Unload (2 gal/CFM)
0%	27%
10%	33%
20%	61%
30%	72%
40%	79%
50%	85%
60%	90%
70%	94%
80%	97%
90%	100%
100%	100%

3.1.3 CFM-bin Hour Profile Analysis Approach

The above methods for determining the instantaneous demand of an air compressor at a given load can be repeated for many bins of hour-CFM operation. This is commonly referred to as a CFM demand profile. A demand profile must be developed to provide accurate estimates of annual energy consumption. A demand profile typically consists of a CFM-bin hour table summarizing hours of usage under all common loading conditions throughout a given year.

Table 4 provides an example of a compressed air CFM-bin hour profile based on the following assumptions:

- The base-case compressor system consisted of a 75 hp rotary screw compressor with inlet valve modulation (w/blowdown) controls, an adjusted full-load power of approximately 65.5 kW, and a rated flow of approximately 360 SCFM.
- The post-retrofit case compressor system consists of a 75 hp rotary screw compressor with VSD (w/stopping) controls, an adjusted full-load power of approximately 67.5 kW, and a rated flow of approximately 360 SCFM.

The annual CFM profile is used to determine base case and proposed case energy use. For both, compressor electricity demand for each CFM-bin should be determined from actual metering data, spot power measurements, or CFM-to-kW lookup tables. When analyzing metered trend data, the hourly average percentage of power should be used to determine which CFM-bin an individual hour is assigned.

The difference in energy consumption between an air compressor operating in idling mode and being physically shut down can be significant depending on the base case and post-retrofit case methods of system control (as demonstrated by CFM-bin6) where base case consumption includes 13,113 kilowatt hours (kWh) when the inlet valve modulation (w/blowdown) compressor is operating in idling mode for approximately 770 hours per year; whereas the post-retrofit case VSD-controlled system (w/stopping) has zero energy consumption for the same bin-hours. It is also common to differentiate between compressor systems operating in “timed-out” mode versus “shut-down” mode. “Timed-out" mode is generally determined from metering. “Shut-down” mode is typically determined from staff interviews and is verified from metering.

Table 4. Sample Compressed Air CFM-Bin Hour Table Base and Post Cases

CFM-bin #	CFM Load Profile	Base Case: Rotary Screw Compressor With Inlet Valve Modulation (w/Blowdown)				Post Case: VSD Rotary Screw Compressor w/Stopping			
		Percentage of Power	H/Yr	Input Power (kW)	kWh	Percentage of Power	H/Yr	Input Power (kW)	kWh
CFM-bin 1	324	97%	200	63.5	12,707	89%	200	60.1	12,015
CFM-bin 2	288	94%	2,440	61.6	150,231	80%	2,440	54.0	131,760
CFM-bin 3	216	88%	170	57.6	9,799	60%	170	40.5	6,885
CFM-bin 4	180	86%	430	56.3	24,222	53%	430	35.8	15,383
CFM-bin 5	144	82%	1,100	53.7	59,081	41%	1,100	27.7	30,443
CFM-bin 6	0 idling *	26%	770	17.0	13,113	0%	0	0.0	0.0
CFM-bin 7	0 shutdown	0%	3,650	0.0	0.0	0%	4,420	0.0	0.0
Total kWh/yr		**269,153**				**196,486**			

The energy consumption for each CFM-bin is determined from the product of the average compressor demand and the number of hours in each bin (Equation 7). The sum of the kWh bin values gives the annual consumption (Equation 8).

$$\Delta kWh_{bin1} = (\text{Base } kW_{operating\_bin1} - \text{Post } kW_{operating\_bin1}) \times \text{CFM-bin 1 H} \tag{7}$$

$$\Delta kWh_{binN} = (\text{Base } kW_{operating\_binN} - \text{Post } kW_{operating\_binN}) \times \text{CFM-bin N H}$$

where:

Base $kW_{operating\_bin1}$ = baseline demand at part-load associated with CFM-bin 1

Post $kW_{operating\_bin1}$ = post demand at part-load associated with CFM-bin 1

Base $kW_{operating\_binN}$ = baseline demand at part-load associated with CFM-bin N

$\text{Post kW}_{\text{operating\_binN}}$ = post demand at part-load associated with CFM-bin N

Total energy reduction:

$$\text{kWh/yr} = \sum_{1\text{-}n} [\ \Delta\text{kWh}_{\text{bin1}} + \Delta\text{kWh}_{\text{bin2}} + \ldots + \Delta\text{kWh}_{\text{binN}}\ ] \quad (8)$$

where:

$\Delta\text{kWh}_{\text{bin1}}$ = energy reduction for CFM-bin 1

$\Delta\text{kWh}_{\text{binN}}$ = energy reduction for CFM-bin N

Another common practice is to incorporate day-types into the CFM-bin analysis as compressed air demands are often tied to facility operations and production schedules. This approach can be particularly useful when developing 8,760 load shapes and when calculating peak demand savings. Day-type analysis is also beneficial when estimating savings from leak repairs and upgrading compressed air dryers.

The CFM-bins should be carefully developed to be applicable to the facility operation. Enough CFM-bins should be present to adequately characterize the granularity of operations. At a minimum, characterizing each individual shift and variances between day types (e.g., weekdays vs. weekends) is needed. A consistent method that nearly always provides appropriate granularity is the daily profile analysis, which obtains the average hourly profile for each hour of each day of the week.

3.1.4 Addressing Uncertainty

During compressed air energy efficiency project evaluations, a common issue arises from a lack of information about baseline energy consumption and lack of airflow data. In the absence of measured or trended CFM data, parameters such as load profile and operating hours must be developed by the evaluator, based on interviews with on-site facility personnel, reviews of historical operations/production levels, reported operating schedules, and short-term (two weeks or more) individual compressor power recordings.

Another common finding from compressed air program evaluations is the fact that baseline and post-installation energy savings calculations are not normalized to account for changes in facility production levels. Best practice when estimating the energy savings of a project is to develop correlations between, not only energy usage and airflow, but also production whenever possible. This allows the evaluator to select the optimal normalization parameter to improve the accuracy of estimated savings.

One common method is to measure compressor power. The percentage of power can be correlated to percentage of flow using the appropriate compressor curve for the given control type. In this way, a load profile can be developed that can be used to compare the baseline and post systems at equivalent flow.

For systems with load and unload compressors, timing the load/unload cycles can be an effective way of determining percentage of capacity. A load/unload compressor either produces full flow or no flow; thus, the percentage of measured time when the compressor is loaded is equivalent to percentage of capacity.

3.2 Savings Calculations for Compressed-Air Leak Surveys and Repairs

3.2.1 Quantifying the Compressed-Air Leakage

Before a compressed-air leak survey is conducted, a system leak-down test should be performed to estimate the combined loss (CFM) of compressed-air leaks. Leak-down tests are best performed at the air receiver by isolating the receiver from the supply side of the system. The basic procedures for conducting a leak-down test are:

- Estimate the total storage volume of the compressed-air system, receivers, main headers, etc., in cubic feet.
- During nonproduction hours, start the system and allow it to reach normal operating system pressure.
- Turn off all production loads.
- Shut off the compressor(s).
- Allow the system to “leak down” to approximately half the full load pressure (psig) and record the time it takes to reach this point.
- Use the following formula:

$$\text{Leak Flow SCFM (Free Air)} = [(V \times \Delta P)/(\text{Time} \times P_{alt})] \times 1.25 \tag{9}$$

where:

V	=	total storage volume of compressed-air system in cubic feet
ΔP	=	drop in line pressure during leak down test in psig ($P_1 - P_2$)
P_{alt}	=	atmospheric pressure (psia) corrected for local altitude (elevation)
T	=	time it takes for line pressure to drop by 50% from normal system operating pressure (minutes)

The 1.25 multiplier corrects leakage to normal system pressure, allowing for reduced leakage with system pressure falling to 50% of the initial reading.

In many cases, a leak-down test is impractical or critical users must have air at all times. In these instances, flow should be estimated by measuring compressor power and correlating to flow (reference table/methods above). This should be done during a nonproduction period, such as a weekend. During this test, it is important to identify any non-leak users of air. The measured compressor flow should be reduced by the total air use of the non-leak applications to determine the actual leak volume.

Leakage is expressed in terms of the percentage of system capacity. The percentage lost to leakage should be less than 10% in a well-maintained system (Marshall 2013). Poorly maintained systems can have losses as high as 20% to 30%.

3.2.2 Quantifying the Energy Impacts of Compressed Air Leak Repairs

Energy savings resulting from the repair of compressed-air leaks can be significant. The best method for estimating impacts is the CFM-bin approach highlighted in Section 3.1.3. The baseline load profile is developed and simulated to determine baseline energy usage. The upgrade load profile is then generated showing the flow reduction resulting from the leak repair and simulated to give the energy usage post-repair. The difference in energy usage between the baseline and post-energy simulation is the energy use reduction associated with the leak repair.

The full CFM-bin approach is highly accurate, but it can be time consuming and overly complicated for small projects. It also works best when full trend data are available to develop a CFM demand profile. A simplified method, outlined below, closely approximates the CFM-bin approach. This simplified approach is only applicable under these conditions:

- The compressed air system is well-controlled and operates predictably
- The system uses a single compressor to meet variable loads and functions as the trim compressor
- The flow reduction is small enough that the quantity of compressors operating is unchanged (if the flow reduction is significant enough to shut off a compressor, the CFM-bin method must be used).

If the above conditions are met, use the simplified savings algorithm below to estimate the energy savings of a leak repair:

$$\text{kWh Saved} = \text{repaired leak volume} \times kW_{FL}/CFM_{rated} \times \text{Hours} \times \text{CCAF} \tag{10}$$

where:

kWh Saved	= kWh saved per year
repaired leak volume	= rate of air loss from leaks repaired (SCFM)
kW_{FL}	= rated full load kW of the trim air compressor
CFM_{rated}	= rated CFM output of the trim air compressor
Hours	= annual operating hours of the flow reduction (typically the compressed air system operating hours for leak repair measures)
CCAF	= trim compressor control type adjustment factor

The adjustment factor will vary based on the method of system control. Table 5 presents typical adjustment factors for common control strategies. An adjustment factor should be used to ensure that energy savings estimates accurately represent savings. It is common for vendors to use an average measured kW/CFM value, but this frequently results in overestimated savings. The adjustment factors provided in Table 5 were developed using data from the percentage of power versus percentage of capacity curves in Section 3.1.2 (Table 1 & Table 2). Each CCAF value represents the slope of the performance curve when operating within the 40% to 80% capacity range as this is a common operating range for a trim compressor.

Table 5. Recommended Adjustment Factors for Determining Energy Savings from Compressed Air Leak Repairs

Control Method	CCAF
Reciprocating—on/off control	1.00
Reciprocating—load/unload	0.74
Screw – load/unload oil free	0.73
Screw – load/unload 1 gal/CFM	0.43
Screw – load/unload 3 gal/CFM	0.53
Screw – load/unload 5 gal/CFM	0.63
Screw—load/unload 10 gal/CFM	0.73
Screw—inlet modulation	0.30
Screw—inlet modulation w/unloading	0.30
Screw—variable displacement	0.60
Screw—variable speed drive	0.97
Centrifugal Compressors	Varies[a]

[a] Centrifugal part-load performance should be reviewed individually depending on the facility load. Centrifugal compressors have good part-load performance within the throttle range of about 0.86 for IGV and 0.67 for IBV controls. Below the throttle range, a centrifugal compressor simply discharges excess compressed air generated through the blowoff valve; therefore, if the compressor is operating in blowoff, the CCAF would be 0. A value between the throttle range and blowoff CCAF may be applicable depending on the time a specific compressor typically operates within each range of control.

Below is an example calculation of the estimated energy savings resulting from compressed air leak repairs based on the following assumptions:

- Compressed air is supplied to the system by a 75 hp rotary screw compressor with VSD controls, a full-load power of approximately 67.5 kW, and a rated flow of approximately 360 SCFM. The compressor runs 4,160 hours per year. The estimated rate of air loss from leaks repaired is approximately 58 SCFM.

$$\text{kWh Saved} = \text{repaired leak volume} \times \text{kW}_{\text{FL}}/\text{CFM}_{\text{rated}} \times \text{Hours} \times \text{CCAF}$$

Per Table 5, CCAF for “Screw—variable speed drive” = 0.97

$$\text{kWh Saved} = (58\ \text{SCFM}) \times (67.5\ \text{kW} / 360\ \text{SCFM}) \times (4{,}160\ \text{hours}) \times (0.97)$$
$$= 43{,}883\ \text{kWh}$$

The methods shown for the energy impact of repairing leaks can also be applied to other compressed air measures that reduce flow, such as installing high-efficiency air nozzles or installing no-loss condensate drain valves.

3.2.3 Leak Volume Quantification Best Practices

The following basic procedures should be followed when quantifying energy savings resulting from leak repairs:

- Impacts from leaks should be supported with formal documentation. The rated power input to CFM output (air compressor specific power) should be supported by trended system data whenever possible.
- The leakage rate (CFM) from a compressed-air leak can be estimated based on the system line pressure and approximate orifice diameter of the crack or leak identified. Leakage rate is proportional to the square of the measured orifice diameter. Table 6 shows the leakage rates for various line pressures (psig) and leak orifice diameters (inches). Correction factors for well-rounded versus sharp orifice shapes must be applied to the leakage rates to ensure estimates are conservative.

Table 6. Leakage Rates (CFM) for Different Supply Pressures and Approximately Equivalent Orifice Sizes

(DOE 2013)

Pressure (psig)	Orifice Diameter (in.)					
	1/64	1/32	1/16	1/8	1/4	3/8
70	0.29	1.16	4.66	18.62	74.4	167.8
80	0.32	1.26	5.24	20.76	83.1	187.2
90	0.36	1.46	5.72	23.1	92	206.6
100	0.40	1.55	6.31	25.22	100.9	227
125	0.48	1.94	7.66	30.65	122.2	275.5

Values should be multiplied by 0.97 for well-rounded orifices and by 0.61 for sharp orifices (DOE 2013).

- Once leak repair work is complete the combined air loss (CFM) of the logged leaks that were repaired should be summed and compared to the total leakage determined from the preliminary leak-down test. Identifying all leaks in a compressed-air system is nearly impossible, so it is appropriate to allocate a portion of the leak-down test CFM to "undetected leakage." A post-repair leak-down test should also be performed to quantify leak reduction.

4 Measurement and Verification Plan

This protocol describes methods for estimating gross savings from compressed air projects. When choosing an option, consider the following factors:

- The equation variables used to calculate savings
- The uncertainty in the claimed estimates of each parameter
- The cost, complexity, and uncertainty in measuring each variable
- The interactive effects of concurrently implementing multiple compressed-air efficiency measures.

4.1 International Performance Measurement and Verification Protocol Option

The preferred approach for evaluating compressed air electronically commuted motors (ECMs) is International Performance Measurement and Verification Protocol Option A: Retrofit Isolation (Key Parameter Measurement). Options B, C, and D can be used in limited applications, but Option A is the preferred approach. Discussions on the feasibility and applicability of the other approaches are provided below.

4.1.1 Option A: Retrofit Isolation (Key Parameter Measurement)—Preferred Approach

International Performance Measurement and Verification Protocol Option A (Retrofit Isolation Key Parameter Measurement) offers the best approach for measuring the energy consumption of compressed-air system. Option A relies on field measurements of key performance parameters and estimates of key parameters not selected for field measurements. Field measurements are typically collected for compressor load current (amps) or true root mean square (RMS) power (Watts).

Parameters such as airflow, line pressure, compressor specific power, part-load performance, and operating hours are typically determined from a combination of one-time spot measurements, historical production data, manufacturers' specifications, CAGI standard data sheets, and interviews with the customer. Using Option A, the measurement boundary is established on the line side of the power supply feeding the air compressor or VSD.

Interval field measurements of compressor load current (amps) coupled with spot power measurements or true RMS power (Watts) measurements are used to determine the instantaneous operating load of an air compressor and to develop trends of energy consumption over time (minimum metering period of two weeks). Equation 11 is used to convert interval measurements of load current (amps) and one-time spot measurements of line voltage and power factor into operating load ($kW_{operating}$) for three-phase motors.

$$kW_{operating} = \sqrt{3} \times Amps \times Volts_{RMS} \times PF \tag{11}$$

where:

Amps = measured load current

$Volts_{RMS}$ = measured True RMS phase-to-phase voltage

PF = measured power factor

True RMS voltage, load current, and power factor should be measured with the system operating under all common loading conditions. Each “common loading condition” should correlate with an established bin of hour-CFM operation. The derived operating load for each CFM-bin is then inserted into Equation 7 (most commonly as the parameter “Post $kW_{operating\_binN}$”) to determine annual consumption and energy reduction.

4.1.2 Option B: Retrofit Isolation (All Parameter Measurement)

The savings created by compressed air ECMs can be determined using Option B (Retrofit Isolation – All Parameter Measurement); however, the degree of difficulty and costs associated with enhanced measurement and verification will increase. By definition Option B requires “field measurement of all key performance parameters which define the energy use of the ECM-affected system.” This implies that in addition to measuring load current or true RMS power, the evaluator is required to measure airflow (SCFM) and operating hours. Option B also requires pre-retrofit metering before the measure is implemented.

4.1.3 Option C: Whole Facility

Typically, Option C is not applicable because compressed air is generally not more than 10% of a typical facility’s energy consumption.

4.1.4 Option D: Calibrated Simulation

Option D can be used in circumstances where multiple ECMs are concurrently implemented; however, this approach can be cost prohibitive and is less common when evaluating ECMs only affecting compressed air systems.

4.2 Verification Process

In accordance with Option A, the first step of the protocol entails verifying key data collected on typical program application or rebate forms, including information on the baseline compressor system. This typically includes:

- Number of shifts per day, shift-hours per week, weekend hours per week, and estimated total operating hours per year
- Average air demand (SCFM) for each shift
- Baseline equipment use pre- and post-retrofit (lead, trim, or backup compressor)
- Baseline compressor system type (reciprocating, screw oil-less/oil-flooded, two-stage, centrifugal, vane, etc.)
- Baseline compressor system control type (load/no load, inlet modulating dampers, other)
- Baseline compressor system operating pressure (psig) at rated SCFM
- Manufacturer, model number, system type, control method, nominal horsepower, rated SCFM, operating pressure at rated SCFM, and installation date for the new energy-efficient air compressor.

For compressed-air leak survey and repair projects, the following information is also frequently requested:

- Whether the facility currently has a formal compressed-air leak detection program in place
- An estimate of total plant air leakage as a percentage of total use
- Type and model of leak detection instrument used by the trade ally to conduct the survey.

Some of these data can be verified using a desk review of invoices, manufacturer specifications sheets (which are typically required for rebate/incentive payments), compressed-air survey reports, or an on-site audit of a sample of participants to verify the quality of self-reported information. If efficiency and unit capacity are not collected for each participant, program application requirements should be modified to include these important data.

4.3 Data Requirements

The energy use of a compressed-air system is typically governed by plant production levels. The actual recommended metering duration for any given compressed-air project should be established to represent all operating modes of the facility. This period should span two full operating cycles from maximum energy use (e.g., weekday production) to minimum (e.g., weekend nonproduction) to confirm the rate of recurrence in the metered data. This is also done to evaluate the consistency of operations on a cycle-to-cycle basis and avoid circumstances where data collected during a single cycle coincided with abnormal operations. For most non-weather-dependent compressed-air applications, a metering period of one month or less is acceptable.

Sampling intervals of 30 to 60 seconds are recommended, although sampling should occur at a high enough frequency to avoid aliasing errors associated with rapidly fluctuating system demand. In general, the sampling frequency should be at least twice the frequency of events in the system, such as compressor load and unload cycles. In most applications, a sampling interval of 30 to 60 seconds satisfies this requirement.

The minimum data required to evaluate a high efficiency air compressor replacement project are:

- Equipment manufacturer, model, and serial number
- Compressor system type (e.g., reciprocating, oil-flooded rotary screw, centrifugal)
- Prime mover (motor) efficiency
- Rated compressor shaft horsepower (bhp) or rated compressor horsepower and prime mover (motor) load factor
- Rated fully loaded SCFM output
- Rated input power of the compressor in kW over output flow rate in CFM (at rated pressure)
- Annual operating hours of constant speed or modulating compressors at a range of loadings

- Load factor of baseline constant speed or modulating compressor
- Percentage of CFM versus percentage of kW curve of new variable displacement capacity or VSD compressor
- Type of control system (modulation, load/no-load, VSD, variable displacement, etc.).

All of the above listed parameters should be gathered for both the baseline and energy-efficient equipment.

Parameters to be spot-measured during the verification include:

- Integrated true RMS kW three-phase power under all common compressor loading conditions.

Parameters to be metered or trended:

- Preferred method: True poly-phase RMS power (kW): This protocol prefers a trend log of true poly-phase RMS power for the circuit powering the VSD compressor. The selected sampling interval should be at a high enough frequency to avoid aliasing errors and at least twice the frequency of events in the system. In general, a sampling interval of once per minute is preferred.
- Alternative method #1: In lieu of true power metering, trending of current (amperage) combined with several one-time true power measurements can be used for base-loaded/constant speed systems. This method can also be used with variable frequency drive compressors as long as true-RMS current transducers are used.
- Alternative method #2: If independent true power metering or trending of current (amperage) coupled with spot power measurements is not possible, it is acceptable to use trend data from a central master control or building automation system. It is preferable to have building automation system trend logs of true poly-phase RMS power with a maximum sampling interval of once per five minutes, and one minute or less is required for load/unload controls. One-time spot power measurements should be performed to verify the accuracy of the control system values.

Additional data required to evaluate compressed air leak survey and repair projects include:

- Compressed-air system specific power (kW/CFM), including compressors, dryers, and significant end uses over a range of CFM loadings
- Supply and demand side one-line diagram showing all generation equipment and significant end uses
- Presence of intermediate pressure and/or flow controllers
- Delivery pressure
- Historical production data for systems affecting compressed-air consumption (number of products produced, active equipment, etc. as appropriate for facility). Production data

should be collected for both the pre-and post-retrofit measurement period and appropriate production adjustments should be made to the collected data.

Data to be collected and utilized, when available:

- Measured or trended airflow (SCFM) data can be quite advantageous when evaluating compressed-air ECMs; however, this information can be difficult to obtain and is not generally collected unless the existing compressed-air system controls already have the capability. In the absence of measured or trended CFM data, the evaluator must develop parameters such as load profile and operating hours, based on interviews with on-site facility personnel; reviews of historical operations/production levels; reported operating schedules, and short-term (2 weeks or more) individual compressor power recordings.

5 Data Collection Methods

5.1 Metering

The typical metering equipment used to measure and trend the energy consumption of a VSD compressor are:

- Handheld (or portable) power meters to measure true RMS voltage, current, power, and power factor at all common loading conditions.
- Current transducers for measuring load current while metering (preferably with a linearity accuracy of ±1.0% of the reading). Recording amp loggers are acceptable as long as spot measurements of compressor power are performed with a handheld kW meter at various loadings.
- Watt-hour transducers to measure true power (kW) of one, two, or three phases of a system.
- Meter recorders (data loggers) with adequate storage capacity to match logging interval and measurement frequency.

The selected measurement equipment should always be installed on the line side of a VSD compressor, not on the load side. Measurements from the output of a VSD compressor can lead to significant data errors. In the pre- and post-retrofit measurement periods, all regularly operating compressors serving a common system should be logged simultaneously regardless of quantity of compressors. Compressors that are used only for backup purposes do not need to be logged, although it is good practice to do so to validate that the equipment was never used. Often post-retrofit only measurements are taken and the pre-retrofit power profile is estimated using the post-retrofit CFM (from kW to CFM conversions), data from the CAGI data sheet for the baseline air compressor system, and generic control curves from Table 1 for the baseline control method.

5.2 Ultrasonic Leak Detectors for Compressed Air Leak Surveys

An ultrasonic leak detector with a frequency response of 35 to 45 kHz should be used to conduct compressed air leak surveys. It is also beneficial to use a set of noise attenuating headphones designed to block intense sounds that often occur in industrial environments so that the user may easily hear the sounds received by the instrument.

Ultrasonic leak detectors are an effective tool for identifying and locating leaks in a compressed air system, but should not be relied upon for quantifying the rate of leakage. The accuracy of these devices are dependent on operator experience and proximity to source; they are inherently inaccurate as leakage rates are not directly measured and instead are correlated based on the amount of sound produced by a given leak in decibels. The best practice for quantifying rate of leakage is to conduct leak-down tests prior to and immediately following leak repairs to determine the actual system impact.

6 Methodology

6.1 General Discussion

The primary energy savings verification method is to monitor, by metering, energy use over a time period that reflects a full or complete range of the underlying operations within a specific industrial facility. Monitoring for periods of less than 1 year, as is most often the case, will require that annual energy use be approximated based on the results of short-term metering and historical production data.

A common issue encountered during compressed-air energy efficiency project evaluations is a lack of information about baseline energy consumption. In many instances, baseline consumption must be derived based on pre-retrofit production levels, reported equipment performance, as well as equipment and component specifications. Key parameters to be determined include motor efficiencies, load factors, load profiles, operating hours, total system SCFM and compressor efficacies (kW/CFM). Often, this information must be gathered through interviews with the program participant, implementer, or energy advisor directly involved with the project.

Other resources frequently used to inform baseline assumptions include:

- Equipment tags[2]
- Historical trending from an EMS
- Engineering reports and calculations generated during the design and application phases of the project
- Rebate or incentive program application forms.

When determining energy savings for VSD compressors, production data must be normalized to an independent normalizing variable. A unit indicating a relative level of production should be obtained from the site, often provided as units produced, hours of machine operation, or labor hours, depending on the site and the availability of information.

Preferably, the independent variable would be collected with sufficient granularity so a correlation can be developed between the measured compressed air energy consumption and the independent variable. The correlation should have a coefficient of determination (R^2) value of at least 0.90 to be of value to the analysis. The pre- and post-retrofit periods should then be normalized to an annual variable for units of production to determine the annual effect of the system improvement. If an annual value is unavailable, using an average of production between the pre- and post-retrofit periods can be acceptable.

Many sites may not be able to provide an independent variable for normalization. In these cases, normalizing to flow is an acceptable alternative. Two methods are used depending on the type of ECM implemented:

[2] It is common for baseline compressor systems to be salvaged or kept in service and converted to an emergency backup role. This provides an opportunity for the evaluator to observe and collect information from equipment tags.

- ECMs that reduce system flow (leaks, air nozzles, condensate drains): For this type of upgrade, the individual installed components should be inspected and CFM reduction confirmed. The flow reduction can then be modeled via a bin table approach using the measured compressor data and simulating the decrease in energy consumption caused by the decrease in flow.
- ECMs that improve system specific power (new air compressors, compressor controls): For this type of upgrade, the system CFM should be determined at each measured point for both the baseline and the installed systems. The CFM should then be compared. The pre- and post-retrofit periods should be normalized to an annual CFM demand profile. The system should then be simulated via a bin table approach at the normalized CFM level using the correlation between flow and power for the respective system.

In a new construction situation where past process production volume and past energy consumption data are unavailable, the determination of energy use per unit of production will have to be based on some form of comparable site such as a similar process in-house or in-company at another facility. For new construction or normal end-of-life replacement projects the baseline system efficiency is determined from the minimum allowed by current local jurisdictions.

The key parameters from Equation 3 are: % Power, ΔkW, and annual operating hours. Each will fluctuate based on the operating load profile of the VSD compressor. Actual post-retrofit consumption can be determined from the sum of multiple iterations of Equation 3, where a unique calculation must be performed for each common loading condition (i.e., using a bin table method). The compressor load profile dictates the number of iterations. Metering generally provides this information.

6.2 Step-by-Step Procedures for Evaluating High-Efficiency/Variable-Speed Drive Air Compressor Installation Projects

This section of the protocol summarizes the basic step-by-step procedures to be performed when evaluating a high-efficiency/VSD compressor replacing a modulating compressor measure.

Step 1: Collect product performance data for baseline and new high efficiency/VSD air-compressor equipment. If product literature is not available, data should be collected from the equipment nameplate. Product literature may be obtainable online after leaving the site using the manufacturer and model number. A sample data collection form is shown in Table 7. Note that the data fields shown in Table 7 should be collected for both the baseline and new equipment.

Table 7. General On-Site Data Collection Form for Air Compressor

Air Compressor General Data Collection Form			
Manufacturer:		Rated Flow (SCFM):	
Model Number:		Pressure at Rated Flow (psig):	
Nominal hp:		Full Load kW_{rated}:	
Drive Motor Efficiency:		Fan Motor hp and Efficiency (if applicable):	
Air-Cooled/Water-Cooled:	❒ Air-cooled ❒ Water-cooled		
Duty:	❒ Lead (Primary) ❒ Trim (Secondary) ❒ Back-up		
Compressor Type:	❒ Rotary Screw (oil-flooded) ❒ Rotary Screw (oil-less) ❒ Centrifugal ❒ Other______		
Control Type (Screw Compressors)	❒ On/Off		
	❒ Load/Unload Total Storage Volume (gallons): _______		
	❒ Inlet Modulating Dampers ❒ w/blowdown ❒ w/o blowdown		
	❒ Variable Speed Drive (VSD) ❒w/unloading ❒ w/stopping		
	❒ Variable Displacement ❒ Other		

Step 2: Determine compressor power at full load for baseline and new high efficiency/VSD air-compressor units using either CAGI performance sheet data, metered full-load and fully unloaded kW data, or derived using Equations 1 and 2. On projects involving the replacement of an older air-compressor system, the evaluator may encounter some difficulty in locating CAGI data sheets, product literature, or manufacturer specifications for the baseline system. In the absence of historical metering data or product literature, the full-load kW for an air compressor system can be derived using Equation 1 or 2:

$$\text{Full Load } kW_{rated} = \frac{(\text{Compressor hp}) \times LF_{rated} \times (0.746 \text{ kW/hp})}{(\eta_{motor})} \quad (1)$$

$$\text{Full Load } kW_{rated} = \frac{(\text{Compressor hp}) \times LF_{rated} \times (0.746 \text{ kW/hp})}{(\eta_{motor}) \times (\eta_{VSD})} \quad (2)$$

where:

Compressor hp = compressor horsepower, nominal rating of the prime mover (motor)

0.746 = horsepower to kW conversion factor

η_{motor} = motor efficiency (%)

η_{VSD}	= variable-speed drive efficiency (%)
LF_{rated}	= load factor of compressor at full load (typically 1.0 to 1.2)

Typically the compressor hp will be known by the customer or on-site personnel. Motor efficiency and load factor may or may not be known by on-site personnel and may need to be estimated using engineering judgment informed by known parameters such as system type, method of control, and age.

Step 3: Once rated compressor power at full load for the baseline and new high efficiency/VSD air compressor have been determined, correct these values for site-specific conditions using Equation 3 or the "rule-of-thumb" approach (Equation 4). The two primary adjustments that must be made pertain to atmospheric pressure based on site elevation above sea level and actual system discharge pressure (psig).

Preferred Approach

$$\mathrm{kW_{adjusted}} = \text{Full Load } \mathrm{kW_{rated}} \; \frac{\left[\left(\frac{P_{discharge} + P_{alt}}{P_{alt}}\right)^{\frac{0.395}{1.395}} - 1\right]}{\left[\left(\frac{P_{rated} + 14.7}{14.7}\right)^{\frac{0.395}{1.395}} - 1\right]} \tag{3}$$

where:

Full Load kW_{rated}	= full load kW of air compressor at full load capacity and pressure (per CAGI data sheet)
$P_{discharge}$	= actual system discharge pressure (psig)
P_{alt}	= atmospheric pressure based on site elevation above sea level (psia)
P_{rated}	= pressure at rated flow (psig) per CAGI data sheet
14.7	= standard atmospheric conditions (psia) at sea level
(0.395/1.395)	= based on the ratio of specific heat for air at standard atmospheric conditions and isentropic compression with constant specific heats

Alternate "Rule-of-Thumb" Approach for Correcting for Discharge Pressure

Although not the preferred approach, a general rule of thumb for air compressors with a rated pressure capacity of 100 psig is: for every 2 psi increase or decrease in discharge pressure, energy consumption will increase or decrease by approximately 1% at full output flow. A sample calculation is shown below:

$$kW_{adjusted} = \text{Full Load } kW_{rated} \times [1+ (((\underline{P_{rated} - P_{discharge}})/2) \times 0.01)] \tag{4}$$

Step 4: Once the rated compressor power at full load for the baseline and new high efficiency/VSD air-compressor equipment have been adjusted for site-specific conditions, develop a CFM demand profile. A demand profile consists of a CFM-bin hour table, summarizing hours of usage under all common loading conditions throughout a given year for the base and post-retrofit case conditions.

Table 8 provides an example of a Compressed Air CFM-bin Hour Profile. The base and post-retrofit case profiles shown in Table 10 were developed based upon the following assumptions:

- The base-case compressor system consisted of a 75 hp rotary screw compressor with inlet valve modulation (w/blowdown) controls, an adjusted full-load power of approximately 65.5 kW, and a rated flow of approximately 365 SCFM.
- The post-retrofit case compressor system consists of a 75 hp rotary screw compressor with VSD (w/stopping) controls, an adjusted full-load power of approximately 69.2 kW, and a rated flow of approximately 365 SCFM.

Table 8. Example Compressed Air CFM-Bin Hour Table - Base and Post Cases

CFM-bin Number	Air Demand Load Profile (SCFM)	%Capacity[a]	Base Case Hours per Year	Post Case Hours per Year
CFM-bin 1	324	90%	2,640	2,640
CFM-bin 2	288	80%	150	150
CFM-bin 3	216	60%	170	170
CFM-bin 4	180	50%	430	430
CFM-bin 5	144	40%	1,130	1,130
CFM-bin 6	0 idling	26%	770	0
CFM-bin 7	0 shut-down	0%	3,650	4,420
	Total Hours		**8,760**	**8,760**

[a]Percentage of flow (part-load fraction) values were determined assuming a rated output flow of 365 SCFM.

Step 5: Once the base and post-retrofit case CFM demand profiles have been developed, calculate the base case and proposed case energy usage. For both base and post-retrofit cases, compressor electricity demand for each CFM-bin should be determined from actual metering data, spot power measurements, or CFM-to-kW lookup tables (refer to Sections 4.3 and 5.1 for guidance on measurement and verification data requirements and data collection methods).

When actual meter or spot power measurement data are unavailable, the percentage of power at part-load for each CFM-bin is typically determined using the calculated percentage of flow values and generic CFM-to-kW lookup tables (see Table 1 and Table 2 in Section 3.1). Percentage of power is influenced by equipment type and method of control. Percentage of capacity versus percentage of power profiles pertinent to the example project for the base and post-retrofit cases are provided in Table 9.

Table 9. Average Percentage of Power Versus Percentage of Capacity for Base Case and Post Case for Example Project

(Scales and McCulloch 2013)

Percentage of Capacity	Base Case: Rotary Screw w/Inlet Valve Modulation (w/Blowdown)	Percentage of Power for Post Case: VSD Rotary Screw Compressor w/Stopping
0%	26%	0%
10%	40%	12%
20%	54%	24%
30%	62%	33%
40%	82%	41%
50%	86%	53%
60%	88%	60%
70%	92%	71%
80%	94%	80%
90%	97%	89%
100%	100%	100%

Using the percentage of power values from Table 9 and the percentage of capacity values calculated in Step 4, the power at part load (kW) for each CFM-bin is determined using Equation 5:

$$kW_{operating} = kW_{adjusted} \times \%\ Power \qquad (5)$$

where:

$kW_{adjusted}$ = Adjusted full load kW

% Power = percentage of power input (%), ratio of the load that a motor is actually drawing relative to the rated full load.

Note: % Power is not a parameter that can be physically measured.

Revisiting the example problem introduced in Step 4, the part-load power (kW) for each CFM-bin is calculated below and is shown in Table 10.

Table 10. Percentage of Power and Operating Load or Base Case and Post-Retrofit Case for Example Project

CFM-bin Number	CFM Load Profile		Base Case		Post Case	
		Percentage of Capacity	Percentage of Power	$kW_{operating}$	Percentage of Power	$kW_{operating}$
CFM-bin 1	324	90%	97%	63.5	89%	60.1
CFM-bin 2	288	80%	94%	61.6	80%	54.0
CFM-bin 3	216	60%	88%	57.6	60%	40.5
CFM-bin 4	180	50%	86%	56.3	53%	35.8
CFM-bin 5	144	40%	82%	53.7	41%	27.7
CFM-bin 6	0 idling	0%	26%	17.0	0%	0.0
CFM-bin 7	0 shutdown	0%	0%	0.0	0%	0.0

Obtaining an actual percentage of power versus percentage of capacity performance curve for the specific air compressor system being evaluated is recommended (if available). A system-specific curve can also sometimes be developed based on information provided on CAGI data sheets. The data presented in Tables 1 and 7 within this protocol could also be used to chart percentage of power versus percentage of capacity in a spreadsheet platform (MS Excel) and develop polynomial fit curves to better estimate part-load values as opposed to using lookup tables.

Step 6: Once the percentage of power and operating load for each CFM-bin have been determined, calculate the corresponding energy consumption using the product of the average compressor demand and the number of hours in each bin for the base and post cases (Equation 7). The sum of the kWh bin values gives the annual consumption (Equation 8).

$$\Delta kWh_{binN} = (\text{Base } kW_{operating\_binN} - \text{Post } kW_{operating\_binN}) \times \text{CFM-bin N H} \tag{7}$$

where:

Base $kW_{operating\_binN}$ = baseline demand at part-load associated with CFM-bin N

Post $kW_{operating\_binN}$ = post demand at part-load associated with CFM-bin N

Total Energy Reduction:

$$kWh/yr = \sum [\ \Delta kWh_{bin1} + \Delta kWh_{bin2} + \ldots + \Delta kWh_{binN}\ ] \tag{8}$$

where:

ΔkWh_{bin1} = energy reduction for CFM-bin 1

ΔkWh_{binN} = energy reduction for CFM-bin N

Using the data from our example project (summarized in Table 11) and Equation 7, the CFM-bin level energy reduction for each bin would be as follows:

$$\Delta kWh_{bin1} = (63.5\ kW - 60.1\ kW) \times 200\ h = 692\ kWh$$

$$\Delta kWh_{bin2} = (61.6\ kW - 54.0\ kW) \times 2{,}440\ h = 18{,}471\ kWh$$

$\Delta kWh_{bin3} = (57.6\ kW - 40.5\ kW) \times 170\ h \qquad = 2{,}914\ kWh$

$\Delta kWh_{bin4} = (56.3\ kW - 35.8\ kW) \times 430\ h \qquad = 8{,}839\ kWh$

$\Delta kWh_{bin5} = (53.7\ kW - 27.7\ kW) \times 1{,}100\ h \qquad = 28{,}639\ kWh$

$\Delta kWh_{bin6} = (17.0\ kW - 0.0\ kW) \times 770\ h \qquad = 13{,}090\ kWh$

$\Delta kWh_{bin7} = (0.0\ kW - 0.0\ kW) \times 4{,}420\ h \qquad = 0\ kWh$

Table 11. Example Project Compressed-Air CFM-Bin Hour Table and Consumption - Base and Post-Retrofit Cases

		Base Case: Rotary Screw Compressor with Inlet Valve Modulation (w/Blowdown)				Post Case: VSD Rotary Screw Compressor w/Stopping			
CFM-bin #	**CFM Load Profile**	**Percentage of Power**	**H/Yr**	**Input Power (kW)**	**kWh**	**Percentage of Power**	**H/Yr**	**Input Power (kW)**	**kWh**
CFM-bin 1	324	97%	200	63.5	12,707	89%	200	60.1	12,015
CFM-bin 2	288	94%	2,440	61.6	150,231	80%	2,440	54.0	131,760
CFM-bin 3	216	88%	170	57.6	9,799	60%	170	40.5	6,885
CFM-bin 4	180	86%	430	56.3	24,222	53%	430	35.8	15,383
CFM-bin 5	144	82%	1,100	53.7	59,081	41%	1,100	27.7	30,443
CFM-bin 6	0 idling	26%	770	17.0	13,090	0%	0	0	0
CFM-bin 7	0 shutdown	0%	3,650	0.0	0.0	0%	4,420	0.0	0.0
Total kWh/yr					**269,153**				**196,486**

Using Equation 7 the Total Energy Reduction resulting from the example project would be:

Total Energy Reduction (kWh/yr) =

$= \sum_{0\text{-}7} [\Delta kWh_{bin1} + \Delta kWh_{bin2} + \Delta kWh_{bin3} + \Delta kWh_{bin4} + \Delta kWh_{bin5} + \Delta kWh_{bin6}]$

$= \sum_{0\text{-}7} [692 + 18{,}471 + 2{,}914 + 8{,}839 + 28{,}639 + 13{,}090 + 0]\ kWh$

$= 72{,}644\ kWh$

7 Sample Design

See Chapter 11: Sample Design for guidance on designing samples to evaluate a program.

Confidence and precision levels are typically determined by specific regulatory or program administrator requirements. In most jurisdictions, evaluation samples should be designed to estimate operating hours and load profiles with a sampling precision of ±10% at the 90% confidence interval.

In addition to sampling errors, errors in measurement and modeling can also occur. In general, these measurement errors are lower than the sampling error; thus, sample sizes are commonly designed to meet sampling precision levels alone.

Sample sizes for achieving the required precision should be determined by estimating the coefficient of variation. These generally range from 0.5 to 1.06 for compressed-air measures, with lower values for more homogeneous populations.

7.1 Program Evaluation Elements

To ensure the validity of data collected, establish procedures at the beginning of the study to address the following issues:

- Quality of an acceptable regression curve fit (based on R^2, missing data, etc.).
- Procedures for filling in limited amounts of missing data.
- Meter failure (the minimum amount of data from a site required for analysis).
- High and low data limits (based on meter sensitivity, malfunction, etc.).
- Units to be metered not operational during the site visit; for example, determine whether this should be brought to the owner's attention or whether the unit should be metered as is.
- Units to be metered malfunction during the mid-metering period and have (or have not) been repaired at the customer's instigation.

An additional 10% of the number of sites or units should be put into the sample to account for data attrition.

At the beginning of each study, determine whether metering efforts should capture short-term measure persistence. That is, decide how the metering study should capture the impacts of nonoperational rebated equipment (due to malfunction, equipment never installed, etc.). For nonoperational equipment, these could be treated as equipment with zero operating hours, or a separate assessment of the in-service rate could be conducted.

7.2 Net-to-Gross Estimation

The cross-cutting chapter, *Estimating Net Savings – Common Practices*, discusses various approaches for determining net program impacts.

8 Looking Forward

VSD air compressor incentive offerings may become less common in the future as regional and state energy codes and standards begin to adopt minimum efficiency requirements similar to those already in effect in California via *Title 24 Energy Efficiency Standards for Residential and Nonresidential Buildings*. The 2013 version of *Title 24* requires that every newly installed compressed air system larger than 25 hp be equipped with at least one trim compressor that is efficient at part loads (i.e. has VSD control) and that compressed air systems with more than one compressor, and a combined capacity of greater than 100 hp, be equipped with a master controller that is capable of determining the most energy efficient combination of compressors to operate within the system based on current air demands.

However, VSD air compressors still remain a popular measure offering amongst commercial and industrial DSM programs and will continue to offer significant savings potential in most jurisdictions for the foreseeable future.

Measurement and Verification Studies

The following evaluations are examples of studies that utilize the methodologies described in this protocol:

- Impact Evaluation of National Grid's 2014 Rhode Island Prescriptive Compressed Air Installations (DNV GL 2016)
- ComEd's Industrial Comprehensive Systems Studies Program – Implementation Contract – Nexant, Inc.
- Duke Energy Non-Residential Custom Program Impact Evaluation.

References

DOE (2013). Compressed Air Tip Sheet #3, "Minimize Compressed Air Leaks." Compressed Air Challenge – Air Best Practices.com. www1.eere.energy.gov/manufacturing/tech_assistance/pdfs/compressed_air3.pdf

DOE (2012). Motor Systems Tip Sheet #11, Adjustable Speed Drive Part-Load Efficiency. Accessed June 2014. www1.eere.energy.gov/manufacturing/tech_assistance/pdfs/motor_tip_sheet11.pdf.

DNV GL. National Grid: Impact Evaluation of 2014 RI Prescriptive Compressed Air Installations. July 2016. http://www.ripuc.org/eventsactions/docket/4654-Pres_CAIR_FinalReport.pdf

Marshall, R. (2013). "Finding and Fixing Leaks." Compressed Air Challenge – Air Best Practices.com. www.airbestpractices.com/system-assessments/leaks/finding-and-fixing-leaks

Scales, W. and McCulloch, D. M. "Appendix 2.A.2." Best Practices for Compressed Air Systems. 2nd ed. Alexandria, VA: Compressed Air Challenge, 2013. 123. Print.

Resources

U.S. Department of Energy – Energy Efficiency and Renewable Energy Department – Compressed Air Challenge. "Improving Compressed Air System Performance: A Sourcebook for Industry." November 2003. www1.eere.energy.gov/manufacturing/tech_assistance/pdfs/compressed_air_sourcebook.pdf

Kleinhenz, P., Seryak, J., Sever, F., Brown, S., and Go Sustainable Energy, LLC. "Comparison of Metering and Verification Methodologies of Compressed Air Systems for Utility-Based Energy-Efficiency Programs: A Case-Study." ACEEE Summer Study on Energy Efficiency in Industry. 2013.

Compressed Air Challenge – Air Best Practices.com. https://www.compressedairchallenge.org/

Compressed Air Challenge Internet Resources

Marshall, R. "VSD Compressor Control." March 2011. Compressed Air Challenge – Air Best Practices.com. www.compressedairchallenge.org/library/articles/2011-03-CABP.pdf

Marshall, R. "Innovative Dual Pressure Control Improves Efficiency." August 2013. Compressed Air Challenge – Air Best Practices.com. www.compressedairchallenge.org/library/articles/2013-08-CABP.pdf

Smith, D.E. "Eliminating Pressure Problems in Compressed Air Systems." April 2012. Compressed Air Challenge – Air Best Practices.com. www.airbestpractices.com/system-assessments/pressure/eliminating-pressure-problems-compressed-air-systems

van Ormer, H. "Drive Down Your Energy Costs with Heat of Compression Recovery." 2013. Plant Services.com. www.plantservices.com/articles/2013/03-heat-of-compression-recovery.html?page=full

Perry, W. "Optimizing the Supply Side with Intelligent Controls." August 2012. Compressed Air Challenge – Air Best Practices.com. www.compressedairchallenge.org/library/articles/2012-08-CABP.pdf

Moskowitz, F. "Heat Recovery and Compressed Air Systems." September 2010. Compressed Air Challenge – Air Best Practices.com. www.compressedairchallenge.org/library/articles/2010-09-CABP.pdf

Wilson, B. "Matching the Supply and Demand in a Compressed Air System." April 2011. Compressed Air Challenge – Air Best Practices.com. www.compressedairchallenge.org/library/articles/2011-04-CABP.pdf

Fox, T.J. "Cycling Refrigerated Air Dryers – Are Savings Significant?" November 2011. Compressed Air Challenge – Air Best Practices.com. www.compressedairchallenge.org/library/articles/2011-11-CABP.pdf

Marshall, R. "How's the Weather in Your Pipes?" November 2012. Compressed Air Challenge – Air Best Practices.com.www.compressedairchallenge.org/library/articles/2012-11-CABP.pdf

Scales, B. and Marshall, R. "Baseline and Monitor Efficiency of Compressed Air Systems." July 2010. PlantServices.com. www.plantservices.com/articles/2010/07MonitorCompressorEfficiency.html

Scales, William, and McCulloch, D.M. (2013). *Best Practices for Compressed Air Systems*. 2nd ed. Alexandria, VA: Compressed Air Challenge, Appendix 2.A.2.

Marshall, R. "Finding and Fixing Leaks" 2013. Compressed Air Challenge – Air Best Practices.com. www.airbestpractices.com/system-assessments/leaks/finding-and-fixing-leaks

Chapter 23: Combined Heat and Power Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in August 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in November 2016.

George Simons and Stephan Barsun
Itron
Davis, California

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68579
October 2017

Contract No. DE-AC36-08GO28308

NATIONAL RENEWABLE ENERGY LABORATORY

Chapter 23: Combined Heat and Power Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance September 2011 – September 2016

This document was republished in August 2017 after a thorough review; no substantive changes were made. This supersedes the version originally published in November 2016.

George Simons and Stephan Barsun
Itron
Davis, California

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68579
October 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The chapter author wishes to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Hugh Henderson of CDH Energy Corp.
- Roger Hill of Navigant Consulting
- Jon Maxwell of Energy & Resource Solutions, Inc.
- William Marin of Itron.

In addition, the chapter author wishes to acknowledge helpful comments submitted through the Stakeholder Review process from:

- Neeharika Naik-Dhungel of U.S. Environmental Protection Agency
- Dana Levy of the New York State Energy Research & Development Authority
- Sue Haselhorst of Energy & Resource Solutions, Inc.
- Bruce Hedman of the Institute for Industrial Productivity.

Suggested Citation

Simons, G.; Barsun, S. (2017). *Chapter 23: Combined Heat and Power Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures*. Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68579. http://www.nrel.gov/docs/fy18osti/68579.pdf

Acronyms

Btu	British thermal unit
CHP	combined heat and power
COP	coefficient of performance
EPA	U.S. Environmental Protection Agency
DOE	U.S. Department of Energy
gpm	gallons per minute
HHV	higher heating value
kW	kilowatt
kWh	kilowatt-hour
LHV	lower heating value
MBtu	thousands of Btu
MMBtu	millions of Btu (thousands of MBtu)
MW	megawatt
NYSERDA	New York State Energy Research and Development Authority
ORC	Organic Rankine Cycle
RMS	root mean square
SGIP	Self-Generation Incentive Program
UHRR	useful heat-recovery rate
UMP	Uniform Methods Project

Table of Contents

List of Figures

List of Tables

1 Measure Description

The main focus of most evaluations is to determine the energy-savings impacts of the installed measure. This protocol defines a combined heat and power (CHP) measure as a system that sequentially generates both electrical energy and useful thermal energy[1] from one fuel source at a host customer's facility or residence. This protocol is aimed primarily at regulators and administrators of ratepayer-funded CHP programs; however, project developers may find the protocol useful to understand how CHP projects are evaluated.

1.1 Scope of the Protocol

The protocol provides a comprehensive method for estimating energy impacts from CHP systems at the customer side of the meter. The protocol's focus on "site energy" rather than "source energy" is consistent with the scope and other protocols developed for the Uniform Methods Project (UMP). Stakeholders may calculate additional metrics, such as source energy impacts or emissions impacts, based on the site energy impacts described in this protocol.

This protocol focuses on CHP systems that are used to meet on-site energy needs and generally sized at less than 5 MW in rated electrical generating capacity. This size range represents 90% of the CHP systems installed since 2000 based on data from the U.S. Department of Energy's (DOE's) CHP Installation Database (DOE 2015).

In addition to providing ways to estimate electricity impacts, the protocol includes algorithms and techniques for assessing CHP fuel impacts and calculating several performance metrics for installed CHP systems. The protocol also allows for the evaluation of different fuel types through the use of energy content for the different fuels. Not every evaluation will need to estimate these performance metrics. In addition, some evaluations may lack data needed to conduct more in-depth evaluations.[2] When such data are missing, the protocol provides default values that can be used to develop impact estimates.

To assist evaluators, the protocol also provides a table to help determine the level of rigor and which equations should be used in estimating impacts. Evaluators should adopt the level of rigor that matches particular evaluation needs and the available data.[3] For larger CHP systems (e.g., 500 kW and more), we strongly urge the use of metered data. In addition, care should be taken to ensure that metered data represents the net electricity generated by the CHP system (net of parasitic loads) and the useful thermal energy actually provided from the CHP system and used by the host site.

---

[1] *Useful thermal energy* refers to thermal energy that is recovered from the CHP system and used to displace thermal energy loads at a host site. Not all heat output from the prime mover can be assumed to be useful heat. Because thermal energy loads can vary, thermal energy available from the CHP system may sometimes exceed the thermal load at the site.

[2] For example, we show methods for calculating hourly impacts that are necessary in evaluating hourly peak demand; however, not all evaluations need to examine hourly impacts and can instead examine only annual energy impacts.

[3] As discussed in the section "Considering Resource Constraints" in the Introduction to this UMP report, small utilities (as defined under the U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol; therefore, alternative methodologies should be considered for such utilities.

For the purposes of this protocol and to ensure consistency with other UMP protocols, we use the following definitions in discussing gross and net electricity:

- **Gross generation** means the electricity produced by the CHP system (not all of which is usable at the host customer site).
- **Net generation** is the gross generation minus parasitic losses. (This is what most evaluators will measure.)
- **Net electricity impacts** means net generation plus any offset chiller energy.
- **On-site net electricity impacts** means net generation plus offset chiller energy minus exported electricity.

To avoid confusion regarding the impacts that can be attributed to the CHP projects in the evaluation, we refer to “net attributable” impacts. Net attributable impacts refer to the net impacts that are separate from the impacts due to free ridership or spillover. Net attributable impacts are considered in Section 6.3, “Net-to-Gross Estimation.”

1.2 Topics Not Covered By This Protocol

The primary focus of this protocol is in estimating energy impacts on the customer side of the meter from installed CHP systems. It is beyond the scope of this protocol to examine the energy impacts at the source of the energy supply (beyond the customer boundary) or the environmental impacts (e.g., greenhouse gas emissions or criteria air pollutant emissions) resulting from CHP systems. Similarly, although CHP systems are a valuable component of the electricity system, it is also beyond the scope of this protocol to provide a means for calculating net electricity system efficiencies or examining the system-wide benefits such as improved reliability or resiliency that CHP may provide to the grid. Because environmental and system-wide electricity impacts can result from a wide variety of energy measures and not only CHP systems, it is appropriate to treat these impacts through a crosscutting protocol.

This protocol is not intended for CHP systems larger than 5 MW.[4] In addition, this protocol does not include an evaluation of bottoming cycles other than those related to steam Rankine cycles.[5]

1.3 Overview of CHP System Applications

For decades, CHP systems sized at 20 MW and more have been widely used in the steel, chemical, paper, and petroleum-refining industries. More recently, smaller CHP systems sized to help meet customer energy needs are being deployed at university campuses, in the food and health industries, and at commercial buildings.

[4] Due to the higher investment associated with these larger systems, we have assumed that the utility or program administrator has worked closely with the CHP project developer and has a good understanding of the project impacts.

[5] Other than the steam Rankine cycle, in this protocol we do not address bottoming cycle CHP technologies such as Organic Rankine Cycle (ORC) because few of these systems appear to be installed through utility programs. The U.S. Environmental Protection Agency’s (EPA’s) market assessment shows that less than 40 ORC-type waste-heat-to-power systems were installed in the United States as of 2012 (EPA and Combined Heat and Power Partnership 2012).

In general, CHP systems are installed to help reduce energy costs by offsetting electricity and other fuel purchases. They achieve these cost savings partly through increased efficiency. Due to the integration of power generation and thermal energy recovery, appropriately designed and implemented CHP systems can be significantly more efficient than separate heat and power generating systems.

Due to their higher overall efficiencies, CHP systems shift electric load away from centralized power plants to the more efficient CHP unit, typically located near the point of use. Figure 1 shows a generalized configuration of a CHP system compared to separate heat and power systems. This figure provides an example of possible differences between separate and CHP systems. Because the local resources powering the grid can vary significantly by location, we strongly recommend using local grid efficiencies and resources for evaluation purposes when possible (EPA 2015).[6,7]

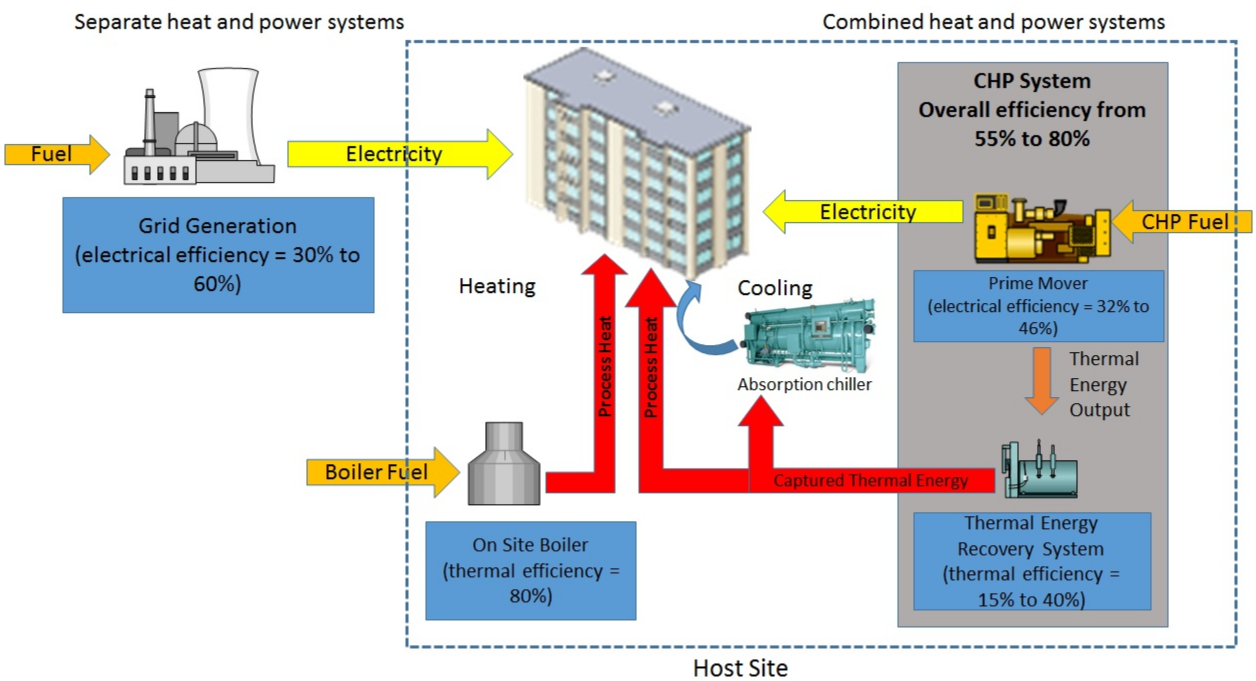


Figure 1. Diagram of separate heat and power compared to CHP

---

[6] Grid generation can occur in a variety of configurations with associated electrical efficiencies. We use a range of central station power plant efficiencies, from 30% to 60% electrical efficiency, as examples. Although we also use a natural gas-fired combined-cycle system in this example, there are instances when a significant portion of the electricity supplied in the local grid comes from coal- or oil-based resources or, conversely, renewable energy resources. Note that when taking into account local renewable energy resources, such as wind or solar photovoltaics, adjustments need to be made to account for the lack of fuel consumption. In addition, line losses associated with the transfer of electricity from the central station system down through the transmission and distribution systems need to be taken into account. See EPA 2015 in "References" section of this document for guidance on calculating fuel and emission savings for CHP systems.

[7] The EPA provides a tool (eGRID) for estimating the electricity resource mix and net generation at various locations throughout the United States. See www.epa.gov/energy/egrid.

Under a separate heat and power system, electricity is provided to the host site from the grid while a boiler, fueled by purchased fuel, provides heat for on-site heat loads. In some instances, heat loads can include absorption chillers to provide on-site cooling needs. In comparison, a CHP system uses purchased fuel to power a prime mover that generates electricity. Thermal energy released from the prime mover is captured in a thermal energy (e.g., heat) recovery system and used to meet on-site heating and absorption cooling loads. The amount of thermal energy recovered *and used* to meet on-site thermal energy needs represents the useful thermal energy.

CHP systems used for self-generation purposes can displace electricity that would otherwise need to be generated and transferred to end uses from electric utilities. Because CHP electricity displacement can often coincide with electric utility system peaks, CHP systems can produce significant peak reduction on the grid.[8] This protocol describes common practice methods to account for hourly and annual energy impacts[9] resulting from installation of CHP systems.

As describe above, CHP systems can supply electricity and thermal energy to a business or industrial plant at a higher efficiency than conventional, separate electricity and thermal generation by capturing much of the heat energy normally wasted in power generation and avoiding line losses. In addition to reducing the total fuel required to provide electricity and thermal energy services to a user, a CHP system may also shift the types of fuel used. Installing a CHP system will generally increase the amount of fuel that is used at the site because additional fuel is required to operate the CHP system compared to the existing boiler that would have otherwise been used to serve the site's thermal demand; however, despite this increase in on-site fuel use, the total fuel use needed to deliver the required electrical and thermal energy services to the facility is reduced by the primary fuel savings generated by the reduced demand from the central station power plant.

Although CHP systems can also affect changes in air pollution emissions, including greenhouse gas emissions, this protocol does not address methods to take into account emission impacts from CHP.

A CHP system consists of a prime mover that consumes fuel to generate electricity and recovers the heat (thermal energy) discharged from the prime mover to produce useful thermal energy. CHP prime movers include a number of different technologies.

[8] In addition, unlike other efficiency measures, CHP systems have the capability to ramp up electricity output, often rapidly. This feature enables CHP systems to be utilized as a dispatchable demand response resource to address local distribution system peak needs even when this does not coincide with the host customer's peak demand. The ability to ramp CHP is dependent on a number of factors, including the ability of the host site to use the captured heat. As more utilities investigate increased integration of distributed energy resources onto the grid, this aspect of CHP systems may become important in future evaluation efforts.

[9] We refer to *impacts* even though other energy-efficiency protocols refer to *savings*. Because CHP projects involve fuel consumption, which may exceed fuel savings, we believe it is more appropriate to refer to energy impacts.

A representative list of CHP prime movers is shown in Table 1.[10] This protocol primarily focuses on natural gas-fueled CHP, but it includes options to estimate energy impacts for CHP fueled by other sources, such as renewable biogas (methane).

Table 1. Representative CHP Prime Movers

Prime Mover	Description	Typical Size Range
Internal Combustion Engine	Reciprocating shaft power can either produce electricity through a generator or drive loads directly. It includes spark ignition and compression ignition engines.	Generally smaller than 5 MW
Gas Turbine	A gas turbine compresses and combusts fuel to create hot gases that are routed into the turbine, spinning the turbine blades. The rotating blades spin a generator to produce electricity.	500 kW to 40 MW
Microturbine	A microturbine is similar to gas turbine in that is uses burner exhaust gases to spin a generator.	30 kW to 250 kW
Fuel Cell	A fuel cell produces an electric current and heat from a chemical reaction between hydrogen and oxygen rather than through combustion.	Generally smaller than 5 MW
Steam Turbine	A steam turbine converts steam energy from a boiler or heat-recovery process into shaft power with a turbine.	50 kW to 250 MW

CHP systems often include auxiliary equipment such as pumps for circulating heat transfer fluids and fans for auxiliary heat rejection. In addition, CHP systems may be connected to other energy processes (e.g., absorption chillers) to help reduce electricity consumption at the host site.

The primary drivers of the electricity and fuel impacts of CHP systems are CHP system efficiencies and utilization:

- **Efficiency**—the effectiveness of fuel conversion and heat recovery in providing electrical and thermal energy services from a CHP system. The two components of overall CHP efficiency are:
    - **Electrical efficiency—**ratio of net electricity generation to fuel consumption[11]
    - **Useful heat-recovery rate (UHRR)—**ratio of heat recovered and used on-site to electricity generation (units: MBtu/kWh or MMBtu/MWh).
- **Utilization**—the extent to which a CHP system is actually used.[12] This performance driver depends on the percentage of time the system is operating as well as on the degree

[10] Other than the steam Rankine cycle, in this protocol we do not address bottoming-cycle CHP technologies such as ORC because few of these systems appear to be installed through utility programs. The EPA's market assessment shows that less than 40 ORC-type waste-heat-to-power systems were installed in the United States as of 2012 (EPA and Combined Heat and Power Partnership 2012).

[11] Note that electrical efficiency is dimensionless by this definition because energy input and energy output are both the same units.

to which the system operates at rated capacity when running. (i.e., actual annual gross kWh generated/system rated kW times 8,760 hours).

Efficiency and utilization are also parameters that can be used in the evaluation in estimating electricity and fuel impacts.

Table 2 lists "target" operational characteristics, such as electrical and overall CHP efficiencies, and UHRR. The targets represent operational characteristics taken from the EPA and Combined Heat and Power Partnership's 2015 *Catalog of CHP Technologies*. The target values represent operations at ideal conditions and are based on a combination of equipment manufacturer specifications and a range of equipment sizes and assumed optimal conditions. For example, the optimal conditions assume that 100% of the thermal energy captured in the heat-recovery system can be used on-site. Evaluators may find observed values can be lower than the EPA targets for several reasons. For example, if evaluated systems are older, the observed values may reflect lower availability due to increased downtime. Similarly, low useful heat recovery rates may reflect there is not a good match between the thermal energy captured by the heat-recovery system and the thermal loads at the host site. We recommend the use of metered data in lieu of assumed values. Although thermal metering represents an additional cost, metering of the amount of thermal energy supplied to the host site (i.e., the useful heat) may be warranted if useful energy recovery is an important factor in the evaluation.

Table 2. Targeted CHP Operational Characteristics[13]

Prime Mover	Electrical Efficiency (HHV)[14]	Overall CHP Efficiency (HHV)	Targeted UHRR (MBtu/MWh)
Internal Combustion Engine	27%–41%	77%–80%	2,996–6,698
Gas Turbine	24%–36%	66%–71%	2,843–6,682
Microturbine	22%–28%	63%–70%	4,265–7,444
Fuel Cell	30%–63%	55%–80%	2,843–5,687
Steam Turbine	5%–40%	near 80%	Not Available

[12] We use capacity factor as "the unrestricted power output of the system divided by the installed capacity" and utilization as "the actual averaged system power output divided by the installed capacity."

[13] The targeted electrical efficiencies and overall CHP efficiencies are from the EPA and Combined Heat and Power Partnership's *Catalog of CHP Technologies* (2015), tables 1–3. The targeted UHRR are calculated based on the electrical and overall system efficiencies.

[14] Higher heating value (HHV) takes into account the latent heat of vaporization of water in the combustion products. Because CHP systems inherently recover some of this heat in the heat-recovery process, we use HHV in reference to efficiencies. In addition, another advantage of using HHV is that it allows for direct comparisons to boilers.

As UHRR increases and offsets on-site boiler fuel, it drives up fuel savings. In turn, the more that useful heat-recovery offsets boiler fuel use during the year, the annual fuel savings tend to decrease.[15] Similarly, the use of prime movers with higher electrical efficiency can result in increased electrical savings through greater displacement of lower efficiency grid-supplied electricity. In this situation, increased utilization of higher electrical efficiency prime movers drives up annual electricity savings.

However, CHP prime movers consume fuel, which affects the overall fuel impacts. Because the prime mover consumes more energy (as fuel) than can be recovered by the heat-recovery system, increased utilization of the CHP system tends to increase annual fuel consumption. Last, thermal energy recovered by the CHP system may be used to drive an absorption chiller to satisfy the cooling load. In this situation, the CHP system offsets the operation of an electric chiller and therefore helps reduce electricity consumption.

The actual performance of individual CHP systems is based on information from input and output energy flows. Typical CHP system components and energy flows are depicted graphically in Figure 2.[16]

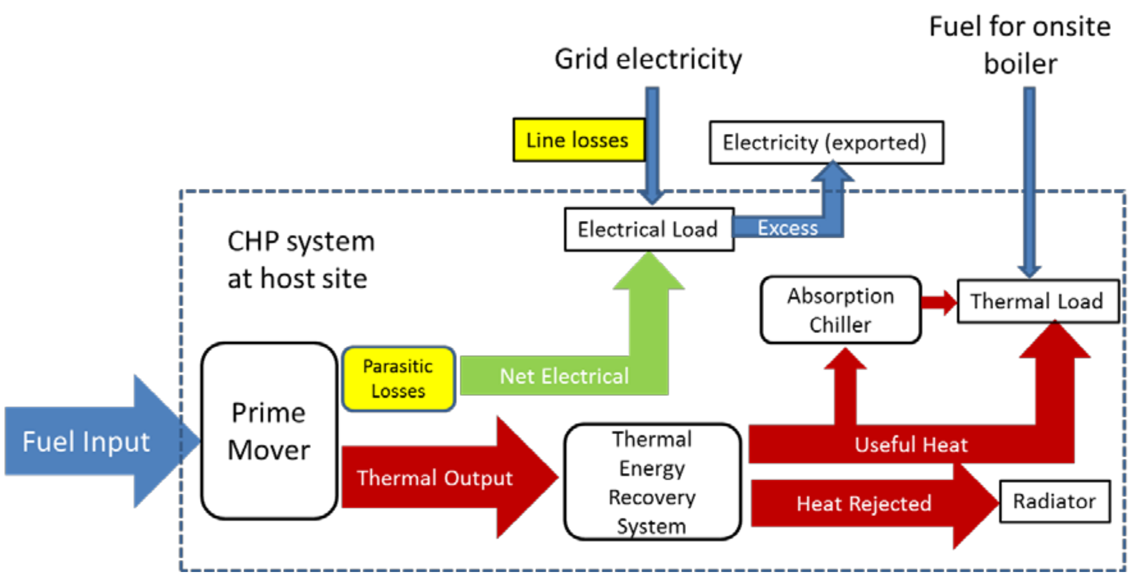


Figure 2. Schematic of CHP component and energy flows

The prime mover consumes fuel to produce gross electricity. Parasitic losses reduce the amount of electricity available for actual use (i.e., net electricity). The net electricity serves on-site electrical loads that would otherwise be served by the grid, thereby reducing grid-generated electricity required by the customer. In certain instances, electricity generated by the CHP

[15] Note that fuel savings is decreasing from the top of the pyramid down; consequently, as the useful heat recovery increases, it pushes the fuel savings upward, thereby increasing fuel savings.
[16] Parasitic losses can occur with a variety of the equipment associated with the CHP system (e.g., pumps and fans for moving fluids or gases). For simplicity's sake, we have only referred to parasitic losses as though they are directly associated with the prime mover.

system may exceed the electrical load of the host site, and, if allowed, the electricity can be exported to the grid.[17] In the course of consuming fuel, thermal energy is generated by the prime mover. A thermal energy (heat) recovery system captures some fraction of the thermal energy generated by the prime mover to serve on-site thermal loads. In some instances, the on-site thermal load may decrease suddenly, and the amount of recovered heat exceeds the on-site load. In those situations, the excess heat is rejected through a "dump radiator." In some instances, useful heat is supplied to an absorption chiller, which can offset electricity normally consumed by an on-site electrical chiller or reduce other electrically served cooling loads. By measuring the amount of fuel consumed by the prime mover and the electricity and useful heat supplied to the host site by the CHP system, we can estimate energy impacts from the system.

[17] Not all utilities allow CHP systems to export electricity to the grid; however, a good example of where this is allowed is under California's Self-Generation Incentive Program (SGIP). Under the SGIP, CHP systems are allowed to export up to 25% of their annual energy demand.

2 Application Conditions of Protocol

Energy-efficiency program administrators may treat CHP systems as a separate and distinct program, or they may include CHP systems as part of a broader population of commercial, multiunit residential, or industrial custom measures.

Energy-efficiency programs that support CHP systems typically provide technical and/or financial assistance to help lower market barriers or help increase customer benefits. Some of these activities may affect the amount of information available for measurement and verification and therefore affect estimated savings. CHP support mechanisms may include the following activities:

- **Prescriptive technology catalogs.** To help reduce costs, accelerate deployment, and increase customer acceptance of CHP systems, program administrators may develop a catalog of standardized sizes, configurations, and installation methods for CHP systems. For example, New York State Energy Research and Development Authority (NYSERDA) uses a prescriptive CHP catalog approach in its CHP Acceleration Program (NYSERDA 2016). Under this approach, programs may the support the installation of only prequalified and conditionally qualified CHP systems by approved CHP system vendors. Typically, these approaches will also include standardized metering installation methods, which can help provide measured performance data on the CHP systems.
- **Training and outreach.** CHP system performance is inherently tied to customer operations and business practices. For example, a business that operates only eight hours per day, 5 days per week and has low thermal energy demand will have lower potential for energy savings from use of CHP than a business that operates 24 hours per day, 7 days per week and has consistently high thermal energy demands. Program administrators may provide training and outreach to educate prospective end users about the "fit" of their business to a CHP project. In addition, program administrators may offer feasibility studies or software tools to help customers better understand CHP project costs and impacts.[18]
- **Rebates or financial incentives.** Program administrators—such as those in California, Massachusetts, and New York—often provide rebates or incentives for customers to install CHP systems that meet specific criteria (e.g., technology type, minimum electrical or system efficiency). Among the types of rebates that can be provided are up-front payments paid per unit of installed capacity (i.e., $/kW) or performance payments paid out per unit of delivered capacity power or energy. In addition, additional "bonus" rebates may be provided to promote the use of special fuels, a higher level of performance, or other preferences (e.g., use of equipment manufactured in the state or use of local installation companies).[19]

[18] For example, utilities participating in the Massachusetts CHP Program require applicants to use a Benefit Cost Model, which takes into account power produced by the CHP system, parasitic losses, quantity and type of fuel consumed, as well as fuel displaced, and timing of power production and thermal loads (Mass Save 2014).
[19] For example, under California's SGIP, CHP systems powered by biogas fuels receive a "biogas adder," whereas CHP systems developed by a California supplier receive additional incentives (Pacific Gas and Electric Company 2015).

- **Demonstrated savings.** The protocol gives guidance for estimating demonstrated savings through actual operation and monitoring. Estimating expected savings from design documents is not supported or recommended with this protocol.

This protocol provides direction on how to evaluate impacts from CHP systems using a consistent approach. The protocol is applicable to new CHP systems and systems that are acting as a retrofit to existing boilers. It does not apply to situations where there was an existing CHP system. This protocol evaluates only installed CHP system impacts. It does not address impacts achieved through training or through market transformation activities.

3 Impact Calculations

This section presents equations for high-level gross impacts that apply to all CHP systems.[20] When evaluating the impacts of CHP systems, electrical, thermal energy, and fuel impacts must be evaluated.[21]

Impacts are all presented on an hourly or finer interval basis.[22] Hourly impacts are summed during the course of the year to calculate annual impacts.[23]

3.1 Determining Electricity Impacts

Note that in some instances CHP projects generate more electricity than can be consumed on-site, and they may be allowed to export electricity to the grid. Because most other energy-efficiency measures do not export electricity, this may be a source of confusion in assessing electricity impacts. For CHP projects, exported electricity should be included in the impacts and noted explicitly. In the following sections, we provide methods for estimating electricity impacts. Although a key priority is the estimation of annual impacts, we provide methods that enable hourly impacts to be estimated. Hourly estimates are important in determining the impacts of CHP systems on utility peak demand. Because peak demand is an hourly occurrence, it requires a method for estimating hourly electricity impacts.

Equation 1a: Hourly net electricity impacts:

$$(Net\ Electricity\ Impacts)_t = [(Gross\ Electricity\ Generated)_t - (Parasitic\ Losses)_t) + (Offset\ Chiller\ Electricity\ Use)_t]$$

where:

$(Gross\ Electricity\ Generated)_t$	=	electrical energy generated at hour t by the CHP equipment; units: kWh
$(Parasitic\ Losses)_t$	=	electrical energy losses at hour t due to pumps, etc., that are required for CHP operation. Ideally, metering would be set up such that any measured generation is the net of parasitic losses, not gross; units: kWh
$(Offset\ Chiller\ Electricity\ Use)_t$	=	electrical energy offset from electrical chillers at hour t if heat from the CHP measure is driving an absorption chiller; units: kWh.

[20] In this instance, we refer to gross electricity impacts to distinguish them from net electricity impacts that account for parasitic losses, offset from electric chiller use.
[21] Because thermal energy impacts both electricity and fuel, these impacts are embedded in these two impact areas.
[22] In many instances, metered electrical data is collected in 15-minute intervals. Interval data can be aggregated to hourly values.
[23] In instances where hourly impacts are not of importance, annual data can be used.

Equation 1b: On-site net hourly electricity impacts:

$$(Onsite\ Net\ Electricty\ Impacts)_t = (Net\ Electricity\ Impacts)_t - (Exported\ Electricity)_t$$

where:

$(Exported\ Electricity)_t$ = net electrical energy generated by the CHP system at hour t that exceeds host site demand.

Note that host site electrical loads may not be known on an hourly basis. In that event, assume that all net electricity generated by the CHP system is consumed at the host site.

Annual net electricity impacts are calculated by summing the hourly impacts for the year.

Equation 2: Annual net electrical impacts:

$$Annual\ Net\ Electricity\ Impacts = \sum_{t=1}^{8760} (Net\ Electricity\ Impacts)_t$$

3.2 Determining Fuel Impacts

Fuel impacts are generally calculated as shown in Equation 3. All energy systems must adhere to thermodynamic laws wherein the amount of energy produced from the system would be less than the energy consumed by the system. As such, CHP fuel impacts are typically negative, meaning that CHP projects consume more fuel to power the prime mover than is saved through recovering the thermal energy from the heat-recovery system, and they offset fuel that would have otherwise been consumed in on-site boilers. Some projects may use one fuel for the CHP system and offset another fuel for heating. For example, a natural gas-fired CHP system may offset an oil-fired boiler. Care should be taken to account for such cross-fuel impacts.

In instances where hourly impacts are deemed unimportant or beyond the scope of the evaluation, the evaluation can use annual fuel data for calculating annual impacts; however, where hourly impacts are important (e.g., in assessing hourly peak impacts, determining efficiency of the CHP system during peak demand, or estimating coincidence between CHP useful thermal energy recovery and CHP generation), hourly fuel impacts need to be assessed. Equation 3 allows for the calculation of hourly fuel impacts.

Equation 3: Hourly fuel impacts:

$$(Fuel\ Impacts)_t = (Fuel\ Offset)_t - (Fuel\ Consumed\ by\ Prime\ Mover)_t$$

where:

$(Fuel\ Offset)_t$ = reduction in on-site fuel consumption at hour t that would have been used for on-site thermal energy needs and is derived exclusively from heat recovered by the CHP system; units: MBtu (HHV basis)

$(Fuel\ Consumed\ by\ Prime\ Mover\ )_t$ = fuel consumed at hour t by the prime mover; units: MBtu (HHV basis).

If there are multiple fuels, fuel impacts are calculated for each fuel type and then summed to estimate total fuel impacts. Note that because fuel consumption is based on an energy (HHV) basis, this equation can be used for multiple fuel types.

If fuel consumption data are not available, the fuel consumption can be estimated based on electrical generation and efficiency, as shown below:

$$(Fuel\ Consumed\ by\ Prime\ Mover\ )_t = \left(\frac{Gross\ Electricity\ Generated_t}{\eta_{EE}}\right)(3{,}412)$$

where:

$\eta_{EE=}$ = electrical efficiency of prime mover (HHV basis)

3,412 = conversion factor 3,412 Btu/kWh.

Section 4.7, "Detailed Procedures," provides more information on determining fuel impacts that take into account electrical efficiency and useful thermal energy recovery.

When multiple fuels are consumed and fuel consumption data are not available, fuel purchase and delivery records should be examined to determine percentage blends of the fuels for each period, *t*. The percentages can then be used to determine fuel impacts.

Annual fuel impacts are calculated by summing the hourly impacts for the year. Again, in instances where hourly fuel impacts are not important, annual fuel data can be substituted. If hourly impacts are important but only annual fuel data are available, hourly fuel rates can be estimated by proportioning them to hourly electricity generation values.

Equation 4: Annual fuel impacts:

$$Annual\ Fuel\ Impacts = \sum_{t=1}^{8760}(Hourly\ Fuel\ Impacts)$$

3.2.1 Special Fuel Situations: Use of On-site and Directed Biogas

Increasingly, CHP systems are being installed in locations such as wastewater treatment plants, landfills, and dairies. In these instances, CHP systems provide benefits by capturing and using the on-site biogas that would have otherwise been vented to the atmosphere or flared. In some of these locations, the host site may use on-site biogas in a boiler to meet on-site thermal needs but not to generate power; consequently, the installation of a CHP system does not increase fuel consumption for on-site biogas applications. For systems fueled by a mix of fuel and on-site biogas, a calculated or measured ratio should be used to calculate the fuel impacts.

Directed biogas refers to biogas that is collected from a landfill, wastewater treatment plant, or dairy facility that may be located far from the facilities that will use the biogas. The procured biogas is processed, cleaned up, and injected into a natural gas pipeline for distribution. There is no requirement that the directed biogas sold to a host site contain a significant amount of the original biogas, and in fact it may contain very little (i.e., molecules) of the original biogas. In this way, directed biogas acts much like a renewable energy credit. The difference is that a natural gas product (i.e., the directed biogas) is sold to customers even though it may contain a very inconsequential amount of actual biogas. For these reasons, directed biogas should be evaluated as having the same energy content as natural gas.

3.3 Determining Energy Offset (Baseline Consumption)

Energy consumed and generated by the CHP system on both an annual and hourly peak basis is relatively simple to calculate from metered data; however, a common challenge in evaluating CHP systems is to identify and determine the baseline energy being offset by the CHP system. In many CHP applications, the CHP system represents the retrofit to an existing boiler; consequently, the on-site boiler fuel consumption represents the thermal energy baseline, which will be offset by CHP thermal energy recovery. In most current situations, CHP systems are designed to match and follow thermal loads of the host site. As a result, it is common to assume that all electricity generated by the CHP system will offset a portion of the on-site electricity loads.

CHP projects may also use recovered heat to drive thermally driven chillers to offset electrical energy that would have been used for cooling. In those instances, baseline chiller electricity demand needs to be taken into account (and can be used to calculate the offset). Likewise, the CHP recovered heat may be used instead of the baseline boiler heat to drive previously operating thermally driven chillers.

Figure 3 shows how the production of electricity and thermal energy from a CHP system can be compared to a baseline.

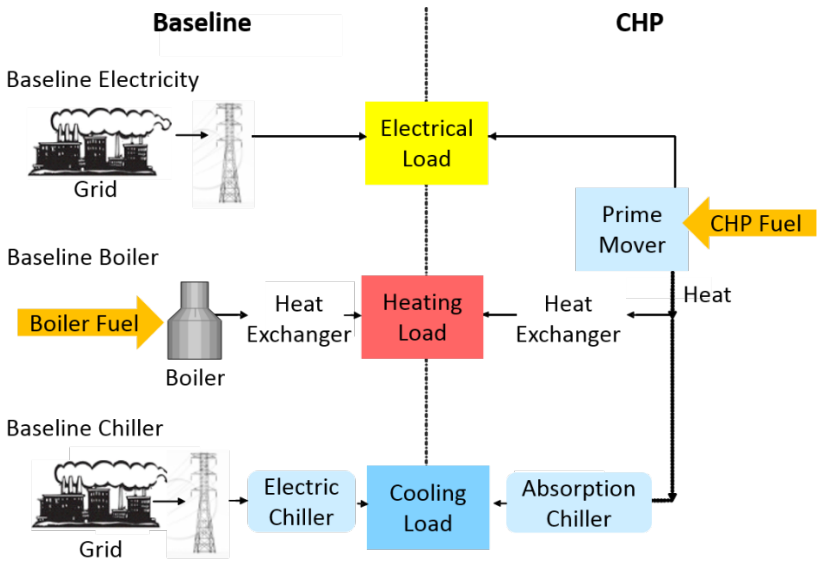


Figure 3. CHP and baseline energy flows

Ideally, site-level data (collected via tracking data or site inspections) are available to identify the boiler, electric chiller, and absorption chiller equipment located at the host site. Although this information may provide equipment specifications, it rarely provides data on operating efficiencies. As a result, some estimates of performance and engineering algorithms are usually required to calculate the amount of boiler fuel displaced by CHP heat recovery and electricity displaced by thermally driven chillers.

Electricity meters should be located such that the metered data explicitly includes the impacts of parasitic loads; however, if this is not the case, parasitic loads must be estimated.[24] The effect of parasitic loads tends to be small (approximately 3% of generation), so assumptions about parasitic loads likely have less of an impact on results than sampling error.[25] Another area that often requires approximation is determining the fraction of recovered heat used to offset heating equipment compared to cooling equipment (when an absorption chiller is present).

If actual on-site equipment details are not available, Table 3 provides recommended default values.

[24] Spot metering can also be used to determine parasitic loads in some instances, but care should be taken to obtain spot measurements at several different operating conditions to determine a reasonable estimate of the parasitic losses. Equipment run time must also be estimated and/or monitored.

[25] Sampling errors occur when CHP systems are looked at in aggregate at the program level.

Table 3. Recommended Default Assumptions[26]

Parameter	Value	Source
Coefficient of performance (COP) for absorption chillers	0.7 for single effect (default) 1.0 for double effect	ASHRAE Standard 90.1-2013, Table 6.8.1C Water Chilling Packages—Efficiency Requirements (full-load)[27]
Electric chiller efficiency[28]	0.6–0.7 kW/t seasonal average or matched by size/type (equal to COP of approximately 5–6)	
Higher heating value of natural gas	1,032 Btu/ft^3	National Energy Technology Laboratory Specification for Selected Feedstock, January 2012, DOE/NETL-341/011812
Heating value of landfill gas	Ranges from 350 to 600 Btu/ft^3 (LHV)	EPA Landfill Methane Outreach Program
Heating value of digester gas	Ranges from 600 to 800 Btu/ft^3 (LHV)	EPA AgStar Program
Boiler efficiency	80%	Rough approximation based on minimum efficiencies specified in ASHRAE Standard 90.1-2010, Table 6.8.1F
Parasitic loads (fan and pump motors, dedicated heating, ventilating, and air-conditioning system and lighting)	3% of generation	Conservative assumption to avoid overstating net electricity, absent spot measurements, or metering
Electrical conversion efficiency	Varies by project and technology (see Table 2)	Project file review, prime mover specification sheet, or average prime mover type efficiencies drawn from industry literature
Fraction of recovered heat used for heat offsets	1.0 if end use of recovered heat is only heating	Approximations if no other data are available. If ex ante analysis includes

[26] Note that lower heating value (LHV) is used for landfill gas and digester gas because this is the most common reference for heating values for these fuels. To convert LHV values to HHV, divide by 0.9.
[27] https://www.ashrae.org/resources--publications/bookstore/standard-90-1. Another source of efficiencies for electric chillers that is broken out by year installed and size is from the Texas Technical Resource Manual (see http://www.texasefficiency.com/images/documents/RegulatoryFilings/DeemedSavings/TRMv3.1v3.docx)
[28] We assume CHP systems are being installed at sites with existing and older chillers (e.g., installed after 2000). Where possible, use ratings specific to the installed chillers.

Parameter	Value	Source
	0.5 if end use of recovered heat is both heating and cooling 0.0 if all recovered heat is used for cooling	division of heat used for cooling vs. heating by season, that division can be reused here.

4 Measurement and Verification Plan

This section contains both recommended approaches to determine CHP energy impacts and the directions on how to use the approaches under the following headings:

- On-Site Inspections
- Vendor and Tracking Data
- Measurement and Verification Method
- CHP Performance Data Collection
- Multiple Fuels
- Interactive Effects
- Detailed Procedures.

4.1 On-Site Inspections

CHP systems installed as part of an energy-efficiency program typically undergo site inspections prior to receiving rebates. Site inspections may be conducted by the evaluation team or by other contractors. Generally, CHP project developers or host site representatives provide pre-inspection data within a program application. On-site inspections are conducted to verify installation of the CHP system nameplate ratings versus tracking data, check gross and net power and/or thermal energy output at the time of the inspection, and collect or coordinate delivery of relevant hourly trend data since the date of "regular" or "normal" operation.

One important aspect of a site inspection may be to establish when the CHP system "entered normal operations." Usually, the date the system enters normal operations is when system commissioning has been completed and the system is considered to be operating much like it will under commercial operations. In some instances, the date at which the system entered normal operations is when incentive checks have been first issued, or it defines the starting point for impact estimates for the program year. Ideally, the threshold for normal operations will have been defined as part of the specific program protocols to avoid confusion.

Site inspection reports should contain:

- Project information (i.e., project name, applicant and host customer name, account number, application number, and facility address)
- Date when the CHP system is considered to have entered normal operations
- Schematic of CHP system (including location of all installed meters) and layout of CHP within host site
- One-line diagrams for electrical distribution and thermal distribution between the prime mover and the useful loads, including rejected energy

- Description of how generated electricity and recovered thermal energy are used at the host site, including identification of the amount of useful thermal energy provided to any absorption chillers to displace electrical loads on electric chillers[29]
- Types of metering being conducted at the site and description of meter download procedures (i.e., how often data is downloaded and to what location)
- Presentation of key trend data, as available.

During the site inspection, the inspector should confirm that the system is a permanent installation connected to the grid and that the generator (prime mover) and heat-recovery system operate as designed.

Table 4 lists representative data collected from site inspections that are important for measurement and verification purposes.

[29] Descriptions of the preexisting operational characteristics of on-site boilers and chillers should be compared to any tracking data obtained for the site.

Table 4. Representative Site Inspection Data

Dates	Fuel Sources	Prime Mover Data	Heat-Recovery System	Absorption Chiller[30]
Inspection date	Primary fuel source (% of energy input)	Technology type	Recovery system type	Chiller type (e.g., single vs. double effect)
Operational date	Flow rate of fuel	Manufacturer	Manufacturer	Manufacturer
	Secondary fuel source (% of energy input)	Model number	Model number	Model number
	Flow rate of secondary fuel	Equipment location	Equipment location	Equipment location
		Prime mover input rate $(MBtu/h)_{HHV}$	End uses served with heat recovery; note whether the BTU meter is net of dumped heat	End uses served with cooling
		Prime mover output (kW)	Hours per year of heat-recovery service	Hours per year of cooling service
		Number of prime mover units	Useful heat-recovery output $(MBtu/h)_{HHV}$	COP
		Total measured power output at inspection (kW); note whether output is net of parasitic loads	Inlet water temperature	Inlet water temperature[31]
			Outlet water temperature	Outlet water temperature
			Water flow rate (gallons per minute [gpm])	Water flowrate (gpm)[32]

[30] Include absorption chiller information in this table only when a new absorption chiller is added as part of the CHP system. Existing absorption chillers are taken into account in the energy offset and through Table 3.

[31] The inlet water temperature to the absorption chiller is the outlet temperature from the heat-recovery system. In general, flows and temperatures for the absorption chiller are not metered unless there is a specific need for this level of rigor. When the evaluation includes numerous CHP projects, it is typical to use the COP to estimate the amount of thermal energy used by the absorption chiller.

4.2 Vendor and Tracking Data

In the course of sizing CHP systems, vendors typically develop estimates of CHP performance, including electricity generation and thermal energy production. In addition, many program administrators require vendors to submit estimated performance, or they may develop their own estimates of CHP performance. Expected CHP performance is contained in "tracking data," which acts as an expected baseline upon which program administrators can project estimated impacts throughout the life of the system. When possible, these vendor or tracking data should be obtained to act as an expected baseline of CHP operation.

4.3 Measurement and Verification Method

This protocol recommends an approach for verifying CHP savings that adheres to Option A—Retrofit Isolation: Key Parameter Measurement—of the International Performance Measurement and Verification Protocol.

Key parameters that require measurement are net electrical generation (and export), useful heat recovery, and fuel consumption. If metered prime mover fuel consumption is not available, it may often be estimated based on prime mover specification sheets and/or data from similar systems. Typically, CHP systems are installed as retrofits, displacing some or all of the thermal output from existing on-site boilers. There is usually no or limited metered data on hourly boiler fuel consumption. This protocol emphasizes metered data collected post-installation (of the CHP system), and it does not include pre-installation data collection requirements.

4.4 CHP Performance Data Collection

To assess energy impacts, data must be collected on CHP performance, including the amount of fuel consumed by the CHP system, electricity generated, and useful thermal energy supplied to the host site. Metered data to be collected include net electricity generated (kWh), net real power delivered (kW), and flow rates and associated inlet and outlet temperatures needed to determine useful thermal energy supplied to the host site. When possible, metered data for fuel consumption of the CHP system should be collected rather than data on site fuel consumption.[33]

When using Option A (the preferred approach) to assess CHP systems, the following measurement and verification elements require particular consideration:

- Measurement Period and Frequency
- Measurement Equipment.

[32] The water flow rate is based on the split between the amount of duty allocated between the heating and cooling loads of the site.

[33] For smaller and older CHP systems, sometimes the only available fuel consumption data is that metered for the entire host site using a utility meter.

4.4.1 Measurement Period and Frequency

Metered data is to be collected post-installation. It is important to use measured data only after the CHP system has completed commissioning and shakedown. The amount of time this takes varies, but measurements can usually start once the CHP system operation approaches "normal" operation (e.g, power and thermal output reach levels that are consistent with expected commercial operation for more than two months). There are two important timing metrics: (1) the measurement periods and (2) the measurement frequency:

- Choose the measurement period (the length of the expected baseline and reporting periods) to capture a full year. This is important in capturing the seasonal impacts of both the CHP system performance and facility operation. If a full year is not available, we recommend capturing at least six months of operational (post-installation) data, with at least one month in summer and one month in winter.
- When hourly impacts are important, choose the measurement frequency (the regularity of the measurements during the measurement period) to provide at least hourly measurements.[34] If an integrating Btu meter is not used, then more frequent data collection intervals may be warranted.

4.4.2 Measurement Equipment

For the key parameters, data may be collected from existing CHP equipment vendor-supplied metering. In the event that the vendor-supplied metering cannot provide enough information,[35] then installing submeters is necessary to obtain data. Use the following guidelines to select the appropriate submetering equipment and procedures[36]:

- Net electricity generation meters:
    - Meters should be located to measure root mean square power output (RMS kW) from the CHP prime mover and ideally after power delivery to all parasitic loads. If not, separate meters or measurements for parasitic loads may be required. Meters should measure net electricity generated (RMS kWh) and net real power delivered (RMS kW).
    - Meters should be capable of collecting data at 15-minute intervals or better and generate accurate date/time stamps for all collected data points.
    - Meters should have the capability to retain collected data in the event of a power outage and should be capable of storing at least seven days of collected data.
    - Meters should have an accuracy of ± 0.5% or meet ANSI C-12.20 certification.

[34] Some CHP incentive programs such as those in California, New York, and Massachusetts are requiring interval meters for measuring electricity generation, useful thermal energy recovered, and fuel consumption. CHP evaluation approaches should take advantage of incentive program metering requirements.
[35] For example, submetering may be required if the existing thermal metering system does not accurately measure useful heat but instead measures only heat output from the prime mover or does not take into account dump radiators. Similarly, some electrical meters may supply only cumulative energy instead of interval energy.
[36] For more on choosing meters, see "Metering Cross-Cutting Protocols" in *Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures* (Mort 2013).

    - Meters can be onboard or external interval data recording meters.
    - When it is feasible within the budget, meters should have the ability to communicate collected data to outside data collection entities (e.g., program administrators).
- Thermal energy recovery meters:
    - Flow meters with "Btu computers" should be insertion-type turbine meters, magnetic flow meters, or ultrasonic flow meters with real-time computation and totalizer.
    - Flow meter/Btu computers should have a field verified accuracy of ±3%.
    - Fluid temperature measurements should be based on temperatures in thermowells or in the flow stream when possible.
    - Flow meters should be calibrated before being placed in the field, verified once installed in the field, and calibrated at least every two years.
    - Metering points should be located to obtain useful thermal energy provided to the host site, taking into account possible radiator dumps.
- CHP fuel consumption meters
    - These are natural gas flow meters with pulse output. Typically, these are rotary-type meters that are temperature and pressure compensated.

Table 5 lists recommended levels of accuracy for the types of metering equipment used for CHP measurement and verification.

Table 5. Recommended Meter Accuracies

Meter Type	Purpose	Accuracy of Meter
BTU meter with flow rates and temperatures	Useful heat recovery	± 3%
Power meter	True RMS power (kW)	± 0.5%
Fuel flow rate meter	Natural gas flow rate	± 1% reading

4.5 Multiple Fuels

Some projects may consume one fuel in the CHP measure to offset a different heating or cooling fuel. For example, the type of fuel consumed by the prime mover may be different than the type of fuel consumed by the existing boiler. Care should be taken to capture all the impacts of the CHP measure on different fuel sources.

4.6 Interactive Effects

For projects evaluated under Option A and that are installed at sites with other efficiency measures, consider how these may interact with the CHP measure. For example:

- A site that installed both a more efficient boiler measure and a CHP system would see no benefits from the new boiler when heating loads were met from the CHP system. In addition, the thermal savings from the CHP system would be reduced somewhat because the boiler efficiency would be higher.

- A site that installed both a CHP system with an absorption chiller and a more efficient electric chiller would get no benefits from the electric chiller when cooling loads are met with the absorption chiller.

4.7 Detailed Procedures

This section presents detailed steps to calculating Equation 1 (electrical impacts) and Equation 3 (fuel impacts).[37] It involves calculating net electrical efficiency as well as electric chiller offset. This section also provides detailed steps to calculating CHP performance metrics such as overall system efficiency and UHRR.

Note that a significant variation in values over time is expected; therefore, each of the equations described in this section should be calculated using the same time frame (annual or hourly). It is not advisable to mix and match time periods when, for example, a one-hour calculation of electrical efficiency is applied to an annual measurement of fuel input.

Some systems may not include all of these parameters, especially absorption chillers, and in rare cases useful heat recovery. The basic components should be directly derived from metered data:

- Electricity generation: directly metered electrical generation, ideally metered as net generation
- Useful heat recovery: directly metered.

4.7.1 Electrical Efficiency

Equation 1 requires knowledge of the electrical efficiency of the CHP system. Electrical efficiency, defined as a measure of how much of the energy in the fuel input is converted to net electricity, is also a key parameter for evaluating CHP performance. This efficiency is largely driven by the type and model of CHP prime mover. Internal combustion engines tend to be more efficient than microturbines, and larger engines tend to be more efficient than smaller engines. Operating conditions also play a role. In general, the closer to full load a prime mover operates, the more efficient the system is at converting fuel to electricity. For larger installations, installing multiple prime movers[38] permits operators to optimize the full loading of each engine.[39] Mathematically, the electrical efficiency is defined as follows:

---

[37]It is typical to calculate electricity impacts first and then fuel impacts because it is usually easier to identify anomalies in electricity output. The electricity impacts can then be used to confirm thermal energy and fuel impacts; however, it is possible to calculate fuel impacts first and then electricity impacts.

[38] When multiple prime movers are used in tandem, the equations should take into account the aggregate capacity of the multiple prime movers; however, if the prime movers are arranged to provide redundancy, care should be taken to aggregate only the systems that will be operated in tandem.

[39] Multiple engines are one simple and effective way of optimizing engine operation to meet varying loads. This method, however, must be balanced with expected load profiles, higher efficiencies often associated with larger engines, and many other factors.

Equation 5: Net electrical efficiency

$$Electrical\ Efficiency_{HHV\ Basis} = \frac{(Net\ Electricity\ Generated_{kWh})}{(Fuel\ Input_{MBtu/hr}) \times \frac{1\ kWh}{3.412\ MBtu}}$$

where:

$Fuel\ Input$ = fuel consumed by the CHP system; make sure to use HHV basis; units: dimensionless.

As noted above, net electrical efficiency requires metered net electricity generation data.

4.7.2 Useful Heat-Recovery Rate

Equation 3 is based on the fuel consumed by the prime mover and the fuel offset. The fuel offset in turn depends on the amount of useful heat recovery achieved by the CHP system. UHRR is one measure of the effectiveness with which thermal energy is recovered from the prime mover and used to meet on-site thermal needs, either on-site heating loads or on-site cooling loads. System design (e.g., sizing) and the timing and magnitudes of facility electrical and thermal loads play key roles in determining a CHP system's heat-recovery rate. Mathematically, the UHRR is defined as follows:

Equation 6: Useful Heat Recovery Rate:

$$Useful\ Heat\ Recovery\ Rate\ (UHRR) = \frac{Useful\ Heat\ Recovered}{Net\ Electricity\ Generated}$$

where:

$Useful\ Heat\ Recovered$ = heat that is actually recovered from the CHP system, including any heat recovered for absorption chiller use and used on-site; units: MBtu (HHV basis).

Note that the UHRR has units of MBtu/kWh.

4.7.3 Overall CHP Efficiency

Electricity generation and recovered heat are combined to form an overall efficiency to quantify how much of the energy input is used. If a CHP system generates substantial quantities of electricity when facility thermal loads are low, large quantities of heat will be rejected to the atmosphere, which will reduce the overall efficiency of the CHP system. Overall efficiency is defined as follows (note the conversions to maintain consistent units):

Equation 6: Overall efficiency:

$$Overall\ Efficiency = \frac{Net\ Electricity\ Generation + Useful\ Heat\ Recovered \times \frac{1\ kWh}{3.412\ MBtu}}{Fuel\ Input\ \times \frac{1\ kWh}{3.412\ MBtu}}$$

Note that, as in Equation 6, useful heat recovery should include any heat recovered for absorption chiller use.

4.7.4 Electric Chiller Offset (Using Thermally Driven Chiller)

Some CHP systems use an absorption chiller to convert useful heat to cooling energy. This allows the CHP system to operate in summer. Equation 8 shows how this electrical cooling offset should be calculated.

Equation 7: Electrical energy offset$_{Chiller}$:

$$ElectricityOffset_{Chiller} =$$

$$\text{Net Electricity Generation} \times UHRR_C \times COP \times \left(EffElecChlr \frac{kWh}{ton-hr\ of\ cooling} \right) \left(\frac{ton-hr\ of\ cooling}{12\ MBtu} \right)$$

where:

$ElectricityOffset_{Chiller}$	=	electricity a power plant would have needed to provide for a baseline electric chiller; units: kWh
$NetElectricityGeneration$	=	net electrical energy generated by the CHP system; units: kWh
$UHRR_C$	=	UHRR that is used to drive an absorption chiller; units: MBtu/kWh
COP	=	COP of the absorption chiller; unitless
$EffElecChlr$	=	efficiency of the baseline electric chiller; units: $\frac{kWh}{Ton-hr\ of\ cooling}$

The hourly impact of CHP systems with a chiller component would be based on the overall concept outlined in Equation 3. It would take into account the boiler efficiency and UHRR and is shown below in Equation 9.

Equation 8: Fuel impacts:

$$Fuel\ Impacts\ = Fuel\ Offset - \ Fuel\ Consumed$$

$$= \frac{Net\ Electricity\ Generation\ \times UHRR_H}{Boiler\ Efficiency} - \frac{Net\ Electricity\ Generation}{Net\ Electrical\ Efficiency} \times \frac{3.412\ MBtu}{1\ kWh}$$

$$= Net\ Electricity\ Generation \times \left[\frac{UHRR_C}{Boiler\ Efficiency} - \frac{1}{Net\ Electrical\ Efficiency} \times \frac{3.412\ MBtu}{1\ kWh}\right]$$

where:

$Fuel\ Offset$	=	reduction in fuel consumption that would have been used for heating that can be attributed to the CHP system; units: MBtu (HHV basis)
$Fuel\ Consumed$	=	fuel consumed by the CHP system. For biogas-fueled CHP systems, this can be zero. This value can be estimated based on electrical generation and efficiency; units: MBtu (HHV basis).
$UHRR_H$	=	URRH that is used to offset on-site heating; units: MBtu/kWh
$Boiler\ Efficiency$	=	efficiency of the boiler of other heating equipment that would serve heating loads in absence of the CHP system; unitless (HHV basis).

4.7.5 Default Assumptions

When possible, the actual efficiencies of heating and cooling equipment should be used in Equation 3 and Equation 8. If this level of detail is not available, Table 3 provides some recommended default assumptions and the reasoning behind them.

4.8 Overall Approach in Estimating Impacts

As identified at the beginning of this protocol, differing levels of rigor can be applied in estimating impacts of CHP projects. Table 6 summarizes the different approaches that can be used in estimating CHP impacts depending on the necessary level of rigor. The rigor and approach can be tailored to the appropriate level of evaluation needs and available data. In addition, Table 6 provides the equations associated with the different CHP performance parameters.

A “full” approach assumes that the evaluation requires not only estimates of energy and fuel impacts but also that these need to be conducted on an hourly basis. For example, this type of approach may be required when the evaluation needs to account for the impact of the CHP systems on peak demand, or if there is a need to determine the degree to which CHP electricity is coincident with useful thermal energy recovery. This approach may typically be used for larger CHP systems or when the CHP systems are part of an incentive program that requires an assessment of peak demand and coincidence of CHP electricity generation to useful thermal energy recovery.

Under a “modified” approach, only electricity impacts are evaluated on an hourly basis, whereas fuel impacts are evaluated on an annual basis. This situation can occur when the evaluation requires an assessment of the impact of CHP on electricity peak demand but there is no requirement to assess the coincidence of CHP electricity with useful thermal energy recovery.

A “simplified” approach is to be used when the evaluation is focused only on annual impacts. This situation may typically be used for very small CHP systems or when CHP systems make up a small portion of an overall energy-efficiency program. Evaluators are warned, however, that if this simplified analysis relies on totalizing meters that report the cumulative usage (total electricity generated, fuel fired, or thermal energy used), additional uncertainty is added to the final results because any meter failures that may have occurred during the aggregation period cannot be detected. This has been a problem particularly with totalizing thermal metering systems.

Table 6. Summary of Approaches for Estimating Impacts

CHP Performance Parameter	Equation(s)	Approach Used for Specified Level of Rigor		
		Simplified	Modified	Full
Net Electrical Impact	1 & 2	Annual	Hourly	Hourly
Net Fuel Impact	3	Annual	Annual	Hourly
Net Electrical Efficiency	4	Annual	Hourly	Hourly
UHRR	5	Annual	Annual	Hourly
Overall Fuel Conversion Efficiency	6	Annual	Annual	Hourly
Electrical Energy Offset	7	Optional	Hourly	Hourly
Fuel Offset	8	Optional	Annual	Hourly

5 Sample Design

At times, evaluators need to assess overall impacts to an energy-efficiency program that has multiple CHP systems. If the number of CHP systems is large, it may be cost prohibitive to collect metered data for all the installed systems. In that event, metered data may be collected from a sample of the operating CHP system.

Consult the UMP's *Chapter 11: Sample Design Cross-Cutting Protocol* for general sampling procedures if the CHP system population is sufficiently large[40] or if the evaluation budget is constrained. Ideally, use stratified sampling to CHP systems by technology and/or the magnitude of claimed (ex ante) project savings. Stratification ensures that evaluators can confidently extrapolate sample findings to the remaining project population. Regulatory or program administrator specifications typically govern the confidence and precision targets, which will influence sample size.

5.1 Detecting and Handling Suspect or Missing Data

Not all received raw metered data are accurate. They may contain errors due to calibration issues, problems with meter operation, or other unforeseen nonsystem issues. All collected data should undergo validation. For example, collected data should be checked to ensure that date/time stamps match actual operation. Similarly, data validation techniques should be used to check and flag suspect data. For example, received electricity generation data that show values significantly higher than those expected given the rated generation capacity of the system should be flagged as suspect. Similarly, data that show zero delivered energy but high values for useful heat recovery should be flagged as suspect.

In some instances, metered data for sites within the sample may not be available for a time period due to outage of the meter or some other nonsystem operational aspect. Ratio estimation is used to generate hourly estimates of performance for periods when observations would otherwise contain missing values.

The premise of ratio estimation is that the performance of unmetered projects can be estimated from similar projects with metered data using a "ratio estimator" and an "auxiliary variable." The ratio estimator is calculated from the metered sample, and the auxiliary variable is used to apply the estimator to the unmetered portion of the data stream. Table 7 provides an example of the different ratio estimators and auxiliary variables used to estimate electricity generation, fuel consumption, or useful heat recovery data.

[40] In general, sampling depends on budgetary considerations; however, a census is recommended at the onset of an energy-efficiency program when CHP systems are beginning to be installed. As the program expands, sampling is recommended when installations of small and same-type systems exceed 20. For larger installations (e.g., 1 MW or larger), energy impacts are significant enough to warrant measurements. In general, sample designs should be set to achieve 90% confidence with 10% precision, depending on budgetary constraints.

Table 7. Example Ratio Estimators and Auxiliary Variables

Variable Estimated	Ratio Estimator	Auxiliary Variable	Stratification
Electricity generation (kWh)	Capacity factor (kWh/kW·hr)	Rebated capacity (kW)	Hourly, by technology type, fuel type, program administrator, operations status, incentive structure, capacity category, and warranty status
Fuel consumption (MBtu)	Electrical conversion efficiency (unitless)	Electricity generated (kWh)	Annual, by technology type
Useful heat recovered (MBtu)	UHRR (MBtu/kWh)	Electricity generated (kWh)	Annual, by technology type

Another issue that arises with collected data is treating "zero" values. In instances when the CHP system is down, a zero value accurately represents nonperformance and should be recorded as a zero value; however, when the CHP system is operational but a zero, null, or missing value is received in the data stream, the zero may simply represent a problem with the metering or the data handling system. Just as validation techniques are used to flag higher-than-expected values, validation techniques should be used to check consistent reporting of missing or bad readings versus true zero values. In the case of suspect useful heat-recovery values, care should be taken to check flow-rate data against temperature data. When data sets contain large amounts of suspect data, it may be necessary to conduct phone surveys to determine the operational status of CHP systems.

6 Other Evaluation Issues

When claiming lifetime and net program CHP measure impacts, consider the following evaluation issues in addition to first-year gross impact findings:

- Early retirement and degradation
- Normalizing CHP performance
- Net-to-gross estimation
- Inter-utility effects.

6.1 Early Retirement and Degradation

CHP projects are often expected to last 10 to 25 years (International Energy Agency 2010); however, during their lifetime, CHP systems can show degradation in availability (which affects capacity factor), electrical, or thermal performance from first-year operations unless a maintenance program is in place. In turn, changes in site operations, fuel, or electricity prices can result in systems being retired after only a handful of years. Evaluators should therefore take care when estimating lifetime performance from first-year savings. That could include persistence studies or leaving metering in place long term to capture savings throughout time. Programs are strongly encouraged to require ongoing metering of electricity output as a requirement for participation.

6.2 Normalizing CHP Performance

The savings from most energy-efficiency measures are correlated to either weather or operating hours; therefore, most energy-efficiency measures can be weather normalized to adjusted weather during the study period to a typical weather period. CHP, however, presents a number of challenges to weather normalization because CHP utilization can be highly variable based on host behavior and other factors. These factors include:

- The cost of fuel (often natural gas)
- The cost of electricity
- The relationship between the cost of fuel and electricity (i.e., if fuel costs rise in relation to electricity, the CHP system will tend to run less; conversely, if fuel costs fall in relation to electricity prices, the CHP system will tend to run more)
- CHP system maintenance (is the system properly maintained on a regular basis so it is available as wanted?)
- Process loads for systems that serve process loads
- Weather for systems that serve heating and cooling loads.

Weather does play a role in CHP operation, but the impact of weather varies from one site to another compared to the other factors listed. CHP host customers can choose to not operate the system and meet their energy needs with more traditional methods. This is quite different than, light-emitting diode lighting, for example, or new space-conditioning equipment that completely replaces the existing equipment so the host can only chose to not have light or heating/cooling or remove the equipment. Therefore, this protocol recommends against attempting to weather normalize CHP performance.

Like other energy measures, CHP performance tends to decrease throughout time, but the impact of this varies and can be influenced by periodic maintenance and servicing. Ultimately, CHP performance should be based on observations (e.g., metered data) that span multiple years. Evaluations that use CHP performance data to normalize operations throughout the life of the system or a program need to account for the factors described above.

6.3 Net-to-Gross Estimation

CHP systems are complex, requiring detailed engineering and sometimes significant effort in obtaining air-pollution control permits and commissioning to bring the system to expected levels of operation. For these reasons, free ridership and spillover do not occur as frequently as they do for other, more common energy-efficiency measures. For some more mature programs, in some instances host sites may install CHP systems without the use of incentives or they may install greater capacity than what can be rebated. As programs mature or as the cost-effectiveness of CHP systems increase, free ridership and spillover need to be taken into account.

The UMP cross-cutting chapter "Estimating Net Savings: Common Practices" discusses various approaches for determining net program impacts. To ensure adjustments to impacts are not double counted at a population level, follow the best practices that include close coordination between (1) staff estimating gross and net impact results and (2) the teams collecting site-specific impact data.

6.4 Inter-Utility and Overall Grid Effects

In some instances, CHP systems may involve multiple utilities. For example, the host site may purchase fuel from one utility and electricity from another. In these situations, evaluators should take care to assess and identify baseline conditions as outlined in Section 3.3, "Determining Energy Offset (Baseline Consumption)." This is particularly important if the impact evaluation baselines are to be used for later CHP cost-effectiveness evaluations.

One of the basic premises of CHP systems is that they offer the potential to provide energy more efficiently and at a lower cost than conventional grid resources. Although defining and providing a means to evaluate net grid impacts is beyond the scope of this protocol, evaluators should make a reasonable attempt to identify the mix of local resources that provide electricity to CHP host sites and the electrical efficiency with which the power is supplied to the site, taking into account transmission and distribution system line losses.

6.5 Other Resources and Examples of Impacts Studies

This protocol provides a methodology for estimating energy impacts from CHP projects that has undergone public review. In developing this protocol, we have relied on a number of past studies that provided insights into the measurement and evaluation of CHP systems. These include the September 2000 measurement and verification guidelines for federal energy projects (DOE 2000), the State of Illinois 2015 *Technical Reference Manual for Combined Heat and Power Systems*, the November 2008 Association of State Energy Research and Technology Transfer Institutions Distributed Generation Combined Heat and Power Long-Term Monitoring Protocol, and the 2005 EPA *Distributed Generation and Combined Heat and Power Field Testing Protocol* (Greenhouse Gas Technology Center Southern Research Institute 2005). These studies served as valuable resources to help augment this protocol.

Impact evaluations are not new to programs incorporating CHP systems. New York has been actively installing CHP systems and NYSERDA has been evaluating their performance for more than a decade. In 2015, NYSERDA released an evaluation report that covers CHP systems installed from 2001 through June 2011 (Energy & Resource Solutions, Inc., and Itron, Inc. 2015). Similarly, Massachusetts and California have been installing numerous CHP systems. Examples of impact evaluations of CHP systems installed in Massachusetts include studies conducted in 2009 and 2010–2011 (KEMA, Inc., Energy & Resource Solutions, Inc., and Itron 2012; KEMA, Inc., 2013). Within California, impact evaluations on CHP systems have been conducted annually since 2003.[41]

[41] Copies of annual impact reports can be downloaded from the CPUC SGIP website at http://www.cpuc.ca.gov/General.aspx?id=7890.

7 References

DOE. 2000. *M&V Guidelines: Measurement and Verification for Federal Energy Projects, Version 2.2* (Technical Report DOE/GO-102000-0960). Washington, D.C.: DOE Office of Energy Efficiency and Renewable Energy. Accessed June 21, 2016. http://www.mass.gov/eea/docs/doer/green-communities/ems/mv-guidelinesv2.pdf.

DOE. 2015. "U.S. DOE CHP Installation Database." Application Version 1.0.10, Data current as of December 31, 2015. https://doe.icfwebservices.com/chpdb/.

Energy & Resource Solutions, Inc., and Itron, Inc. 2015. *Distributed Generation—Combined Heat and Power Impact Evaluation Report (2001–June 2011): Final Report* (Technical Report). North Andover, MA. Accessed June 21, 2016. https://www.nyserda.ny.gov/-/media/Files/Publications/PPSER/Program-Evaluation/2015ContractorReports/2015-Distributed-Genertation-CHP-Impact-Evaluation-Final.pdf.

EPA and Combined Heat and Power Partnership. 2012. "Waste Heat to Power Systems." http://www.heatispower.org/wp-content/uploads/2012/06/EPA-waste_heat_power-report-5.2012.pdf

EPA and Combined Heat and Power Partnership. 2015. *Catalog of CHP Technologies*. Washington, D.C.: IFC International.

EPA. 2015. "Fuel and Carbon Dioxide Emissions Savings Calculation Methodology for Combined Heat and Power Systems." December. https://www.epa.gov/chp/fuel-and-carbon-dioxide-emissions-savings-calculation-methodology-combined-heat-and-power.

Greenhouse Gas Technology Center Southern Research Institute. 2005. *Generic Verification Protocol—Distributed Generation and Combined Heat and Power Field Testing Protocol* (Technical Report SRI/USEPA-GHC-GVP-04). Morrisville, NC. Accessed June 21, 2016. http://nepis.epa.gov/Exe/ZyPDF.cgi/P100CR0O.PDF?Dockey=P100CR0O.PDF.

International Energy Agency. 2010. IEA ETSAP: Technology Brief E04 – May 2010. See http://iea-etsap.org/E-TechDS/PDF/E04-CHP-GS-gct_ADfinal.pdf

KEMA, Inc. 2013. *Massachusetts Combined Heat and Power Program Impact Evaluation 2011–2012* (Technical Report). Burlington, MA: Massachusetts Energy Efficiency Program Administrators and Massachusetts Energy Efficiency Advisory Council. Accessed June 21, 2016. http://ma-eeac.org/wordpress/wp-content/uploads/Combined-Heat-and-Power-2011-12-Program-Evaluation-November-2013.pdf.

KEMA, Inc., Energy & Resource Solutions, Inc., and Itron. 2012. *2010 Combined Heat and Power Impact Evaluation Methodology and Analysis Memo* (Technical Report). Middleton, CT. Massachusetts Energy Efficiency Program Administrators and Massachusetts Energy Efficiency Advisory Council: January 11, 2012.

Mass Save. 2014. *Combined Heat and Power (CHP): A Guide to Submitting CHP Applications for Incentives in Massachusetts*. Accessed June 21, 2016.

http://www.masssave.com/~/media/Files/Business/Applications-and-Rebate-Forms/A-Guide-to-Submitting-CHP-Applications-for-Incentives-in-Massachusetts.pdf.

Mort, Dan. 2013. "Metering Cross-Cutting Protocols." In *Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures* (Subcontract Report NREL/SR-7A30-53827). Golden, CO: National Renewable Energy Laboratory. http://energy.gov/sites/prod/files/2013/11/f5/53827-9.pdf.

National Energy Technology Laboratory. 2012. *Quality Guidelines for Energy System Studies: Specification for Selected Feedstocks* (Technical Report DOE/NETL DOE/NETL-341/011812). Washington, D.C.: U.S. Department of Energy.

NYSERDA. 2016. "NYSERDA PON 2568 CHP Program." Accessed June 21. http://www.nyserda.ny.gov/PON2568.

Pacific Gas and Electric Company. 2015. "2015 SGIP Handbook and Forms." Accessed June 21, 2016. https://www.pge.com/en/mybusiness/save/selfgen/handbook/index.page.

State of Illinois. 2015. *Energy Efficiency Technical Reference Manual: Combined Heat and Power New Measure*. Springfield, IL.

Texas Technical Resource Manual (see http://www.texasefficiency.com/images/documents/RegulatoryFilings/DeemedSavings/TRMv3.1v3.docx).

8 Bibliography

Combined Heat and Power Performance Program Systems Manual, Appendix A. Albany, New York.

Commonwealth of Massachusetts. 2011. *Alternative Energy Portfolio Standards: APS Guideline on the Eligibility and Metering of Combined Heat and Power Projects*. Boston: Executive Office of Energy and Environmental Affairs, Department of Energy Resources. Accessed June 21, 2016. http://www.mass.gov/eea/docs/doer/rps-aps/aps-chp-guidelines-jun14-2011.pdf.

Measurement and Verification Operational Guide: Renewable and Cogeneration Applications. Sydney, Australia: Office of Environment and Heritage. Accessed June 21, 2016. http://www.environment.nsw.gov.au/resources/energyefficiencyindustry/120992reandcogenapp.pdf.

NSW Government. 2012. NYSERDA. 2013. Pacific Gas and Electric Company and the SGIP Working Group. 2015. *2013 SGIP Impact Evaluation*. Davis, CA: Itron. Accessed June 21, 2016. http://www.cpuc.ca.gov/WorkArea/DownloadAsset.aspx?id=7909.

Pacific Gas and Electric Company. 2015. *2014 Self-Generation Incentive Program Handbook*. San Francisco: PG&E Self-Generation Incentive Program. Accessed June 21, 2016. https://www.selfgenca.com/documents/handbook/2014.

Pacific Gas and Electric Company. 2015. *2015 Self-Generation Incentive Program Handbook*. San Francisco: PG&E Self-Generation Incentive Program. Accessed June 21, 2016. https://www.selfgenca.com/documents/handbook/2015.

Shonder, John S. "Measurement and Verification of Savings in Combined Heat and Power Projects" (presentation by Oak Ridge National Laboratory, no publication date). Accessed June 21, 2016. http://gaia.lbl.gov/federal-espc/Summit/DC/DC-Summit-M&V-CHP.pdf.

Chapter 24: Strategic Energy Management (SEM) Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance
July 2016 – April 2018

James Stewart, Ph.D.
The Cadmus Group
Portland, Oregon

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Subcontract Report
NREL/SR-7A40-68316
May 2017

Contract No. DE-AC36-08GO28308

NATIONAL RENEWABLE ENERGY LABORATORY

Chapter 24: Strategic Energy Management (SEM) Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

Created as part of subcontract with period of performance July 2016 – April 2018

James Stewart, Ph.D.
The Cadmus Group
Portland, Oregon

NREL Technical Monitor: Charles Kurnik

Prepared under Subcontract No. LGJ-1-11965-01

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

Subcontract Report
NREL/SR-7A40-68316
May 2017

Contract No. DE-AC36-08GO28308

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at SciTech Connect http:/www.osti.gov/scitech

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
OSTI http://www.osti.gov
Phone: 865.576.8401
Fax: 865.576.5728
Email: reports@osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Road
Alexandria, VA 22312
NTIS http://www.ntis.gov
Phone: 800.553.6847 or 703.605.6000
Fax: 703.605.6900
Email: orders@ntis.gov

Cover Photos by Dennis Schroeder: (left to right) NREL 26173, NREL 18302, NREL 19758, NREL 29642, NREL 19795.

NREL prints on paper that contains recycled content.

Disclaimer

These methods, processes, or best practices (“Practices”) are provided by the National Renewable Energy Laboratory (“NREL”), which is operated by the Alliance for Sustainable Energy LLC (“Alliance”) for the U.S. Department of Energy (the “DOE”).

It is recognized that disclosure of these Practices is provided under the following conditions and warnings: (1) these Practices have been prepared for reference purposes only; (2) these Practices consist of or are based on estimates or assumptions made on a best-efforts basis, based upon present expectations; and (3) these Practices were prepared with existing information and are subject to change without notice.

The user understands that DOE/NREL/ALLIANCE are not obligated to provide the user with any support, consulting, training or assistance of any kind with regard to the use of the Practices or to provide the user with any updates, revisions or new versions thereof. DOE, NREL, and ALLIANCE do not guarantee or endorse any results generated by use of the Practices, and user is entirely responsible for the results and any reliance on the results or the Practices in general.

USER AGREES TO INDEMNIFY DOE/NREL/ALLIANCE AND ITS SUBSIDIARIES, AFFILIATES, OFFICERS, AGENTS, AND EMPLOYEES AGAINST ANY CLAIM OR DEMAND, INCLUDING REASONABLE ATTORNEYS' FEES, RELATED TO USER’S USE OF THE PRACTICES. THE PRACTICES ARE PROVIDED BY DOE/NREL/ALLIANCE "AS IS," AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING BUT NOT LIMITED TO THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL DOE/NREL/ALLIANCE BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER, INCLUDING BUT NOT LIMITED TO CLAIMS ASSOCIATED WITH THE LOSS OF PROFITS, THAT MAY RESULT FROM AN ACTION IN CONTRACT, NEGLIGENCE OR OTHER TORTIOUS CLAIM THAT ARISES OUT OF OR IN CONNECTION WITH THE ACCESS, USE OR PERFORMANCE OF THE PRACTICES.

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the UMP, visit the website, https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at https://energy.gov/sites/prod/files/2015/02/f19/UMPIntro1.pdf.

Acknowledgments

The chapter author wishes to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol:

- Marc Collins of Itron
- Miriam Goldberg, Andrew Styker, and Julia Vetromile of DNV-GL
- Kim Crossman, Erika Kociolek, and Phil Degens of Energy Trust of Oregon
- Deborah Swarts and Bill Biesemeyer of Navigant
- Bill Miller, Aimee McKane, Darren Sholes, Peter Therkelsen, and Shankar Earni of Lawrence Berkeley National Laboratory
- Paul Scheihing and Jay Wrobel of the U.S. Department of Energy
- Jim Volkman of Strategic Energy Group
- Todd Amundson, Lauren Gage, Steve Brooks, and Jennifer Eskil of the Bonneville Power Administration
- Steve Martin of Cascade Energy Engineering
- Bill Koran of SBW Consulting
- Hossein Haeri, M. Sami Khawaja, Jeff Cropp, Jennifer Huckett, Heidi Javanbakht, and Andrew Bernath of Cadmus.

In addition, the chapter author wishes to acknowledge helpful comments submitted through the Stakeholder Review process from:

- Bill Harris of Snohomish Public Utility District
- Jess Burgess of the Consortium for Energy Efficiency
- American Water Works Association.

Suggested Citation

Stewart, James. 2017. *Chapter 24: Strategic Energy Management (SEM) Evaluation Protocol*. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68316. http://www.nrel.gov/docs/fy17osti/68316.pdf

Acronyms

BPA	Bonneville Power Administration
Btu	British thermal unit
CDD	cooling degree day
CEE	Consortium for Energy Efficiency
DOE	U.S. Department of Energy
EM&V	evaluation, measurement, and verification
EnMS	energy management system
HDD	heating degree day
HVAC	heating, ventilation, and air conditioning
IPMVP	International Performance Measurement and Verification Protocol
ISO 50001	International Organization for Standardization (ISO) for an Energy Management System
M&V	measurement and verification
OLS	ordinary least squares
OM&B	operation, maintenance, and behavior
R^2	coefficient of determination
SAS	Statistical Analytics Software
SEM	strategic energy management
SEP	superior energy performance
UMP	Uniform Methods Project

Table of Contents

List of Figures

List of Tables

1 Measure Description

Strategic energy management (SEM) focuses on achieving energy-efficiency improvements through systematic and planned changes in facility operations, maintenance, and behaviors (OM&B) and capital equipment upgrades in large energy-using facilities, including industrial buildings, commercial buildings, and multi-facility organizations such as campuses or communities. Facilities can institute a spectrum of SEM actions, ranging from a simple process for regularly identifying energy-savings actions, to establishing a formal, third-party recognized or certified SEM framework for continuous improvement of energy performance. In general, SEM programs that would be considered part of a utility program will contain a set of energy-reducing goals, principles, and practices emphasizing continuous improvements in energy performance or savings through energy management and an energy management system (EnMS)[1]. An EnMS, as defined by ISO 50001, is a formal process for an organization to establish a policy, objectives, and targets for improved energy performance and to implement and assess energy performance improvement actions taken to meet those objectives and targets. An organization uses this framework to incorporate energy use and consumption into its management processes.

To provide some guidance to utilities in consideration of SEM programs, the Consortium for Energy Efficiency (CEE) has established the following working definition for SEM:

> "Strategic Energy Management can be defined as taking a holistic approach to managing energy use in order to continuously improve energy performance, by achieving persistent energy and cost savings over the long term. It focuses on business practice change from senior management to the shop floor staff, affecting organizational culture to reduce energy waste and improve energy intensity. SEM emphasizes equipping and enabling plant management and staff to impact energy consumption through behavioral and operational change. While SEM does not emphasize a technical or project-centric approach, SEM principles and objectives may support capital project implementation." (CEE 2014a)

The CEE developed a set of three SEM Minimum Elements—customer commitment, planning and implementation, and a measurement and reporting system—supported by 13 specific components of industrial SEM (known as CEE SEM minimum elements) and specific responsibilities for senior managers and the energy management team. It is important to note that not every SEM industrial program incorporates all of these components.

<u>Senior management</u>:

1. Sets and communicates long-range energy performance goals.
2. Ensures SEM initiatives are sufficiently resourced and a responsible individual or team is designated.

[1] As discussed in the section "Considering Resource Constraints" in the Introduction to this UMP report, small utilities (as defined under the U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol; therefore, alternative methodologies should be considered for such utilities.

Designated energy manager or management team:

3. Assesses current energy management practices using a performance scorecard or facilitated energy management assessment.
4. Develops a map of energy use, consumption, and cost, including all significant end-use systems and relevant variables of energy consumption.
5. Establishes clear, measurable metrics and goals for energy performance improvement.
6. Registers or records actions to be undertaken to achieve the energy performance goals.
7. Develops and implements a plan to engage employees in energy performance improvement.
8. Implements planned actions.
9. Periodically reassesses outcomes related to energy performance.
10. Regularly collects performance data to improve understanding of energy use and consumption.
11. Collects and stores performance data related to energy performance improvement metrics and goals, making it available over time.
12. Analyzes energy use and consumption data, determining relevant variables affecting use compared to a baseline.
13. Reports regularly to senior management and others on the results of energy performance improvement actions.

While the CEE developed this list for industrial facilities, the SEM minimum elements also apply to the management of energy use in commercial and institutional buildings, multi-facility organizations, and campus settings.

Currently, many utilities and program administrators offer ratepayer-funded SEM programs that enroll a range of industrial, commercial, and institutional customers (CEE 2016).[2] These utility-administered programs each provide a distinct program design for qualifying participants, which contain some of the CEE elements. Most programs provide participating facilities or organizations with training about energy management practices and EnMS, technical support for implementation, and financial incentives for achieving energy savings, with the objective of integrating SEM into facility or building operations.

Many utility SEM programs expect to save 5% or more of annual facility energy consumption by helping participants to implement these SEM elements (CEE 2014). To acquire savings, utility SEM programs support participants' capability for continuously improving energy performance through the adoption of SEM practices.[3]

[2] CEE (2016) identifies 25-member utilities or program administrators in the United States and Canada that fund industrial SEM programs.
[3] *SEM Program Case Studies Report* (CEE 2015).

1.1 ISO 50001: A Configured Energy Management System (EnMS)

SEM programs fall on a continuum, from those meeting the minimum elements noted above to those that also meet or exceed the requirements of the *ISO 50001 Energy Management System standard*. ISO 50001 is an international standard with a defined "plan-do-check-act" EnMS that sets forth a series of organizational practices to effectively manage energy and continually improve energy performance. ISO 50001 also includes methods for calculating period-over-period changes in energy performance and requires documented evidence of energy performance improvements. Since ISO 50001 is user-administered, organizations seeking ISO 50001 certification are subject to a certification audit conducted by a qualified audit team from a nationally accredited certification body.[4]

An application of an ISO 50001-conformant EnMS is the U.S. Department of Energy's (DOE) Superior Energy Performance® (SEP) certification. SEP builds on ISO 50001 by applying the *Superior Energy Performance Measurement and Verification Protocol* (DOE 2016c) across all energy types to meet specific targets over defined periods of time for measurement and verification of energy performance improvement. In addition, DOE has developed the 50001 Ready program, which follows the *50001 Ready Protocol* (DOE 2017a) and provides DOE (and/or partner) recognition for self-declared conformance to ISO 50001. The 50001 Ready program provides energy and carbon emissions savings calculation and is designed to partner with utilities and other organizations, including state and local governments or multi-facility organizations to support their 'enterprise' of facilities or their supply chain.

1.2 Protocol Objective

The objective of this SEM evaluation protocol is to help program evaluators and administrators accurately assess the gross energy savings of utility SEM programs. This protocol focuses on best practices for estimating energy savings for individual large commercial or industrial facilities, although the protocol also describes methods for conducting analysis to estimate the average savings per facility for a group of facilities.5

As utility SEM programs are a relatively new offering, evaluators are still developing best practices for evaluation. This protocol describes current thinking about best practices; however, it is expected that this protocol will require updating as evaluation approaches improve and consensus builds around the best approaches.

[4] ANSI-ASQ National Accreditation Board. More complete information on ISO 50001 can be found at http://www.energy.gov/eere/amo/iso-50001-frequently-asked-questions

[5] Estimation of average savings for groups of facilities, or "panels" is presented in section 4. For estimation of energy savings from small commercial buildings, see NREL (Agnew 2013).

2 Application Conditions of Protocol

For the purpose of providing guidance about evaluating SEM programs, this protocol differentiates among three categories of SEM programs. The first category includes those that satisfy some or all of the CEE definition of SEM. The second category includes those that require all of the CEE elements and promotes the establishment of an ISO 50001-conformant EnMS. The third category includes those programs that further promote certification to SEP.

This UMP protocol provides guidance for evaluating the savings impacts of SEM programs administered by utilities or other energy efficiency organizations. This protocol applies to all utility SEM programs whether or not they satisfy all of the CEE minimum elements. For utility or energy efficiency organization programs designed to conform with ISO 50001, this protocol incorporates by reference and directs evaluators to use DOE's *Qualified Energy Savings Measurement and Verification Protocol for Industry* (DOE 2017a). For utility or energy-efficiency organization programs designed to conform to SEP, this protocol incorporates by reference and directs evaluators to use the *Superior Energy Performance Measurement and Verification Protocol* (DOE 2017b).

For utility SEM programs that satisfy some or all of the CEE SEM elements, this protocol recommends statistical analysis of metered facility energy consumption for estimating energy savings. A facility is the analysis unit of SEM program impact evaluations and the area over which energy use and consumption will be measured and analyzed. A *facility* may comprise a single building with a single meter or multiple buildings at the same site with multiple energy-use meters.[6] The *reporting period* is when energy savings from SEM activity will be estimated. The *baseline period* is when energy consumption measurements are taken to establish a baseline for the facility's energy consumption.

2.1 Four Key Conditions

Evaluators should apply this protocol when all of the following conditions are satisfied:

- *The evaluation objective is estimating changes in a facility's energy consumption*[7] *(savings) or energy consumption intensity (energy consumption per unit of production output or unit of floor area) from SEM activities*. Estimation of peak demand savings is not covered. While many SEM programs deliver peak demand savings, estimating these savings requires different data and analysis methods from those presented in this protocol.[8]

[6] This definition of a facility will apply to most participants in utility SEM programs; however, some participants such as water utilities and waste water treatment facilities have complex distribution and pumping systems that do not have simple boundaries. Many opportunities for reducing energy consumption through SEM may exist in their distribution networks. The definition of facility is not intended to preclude the participation of water utilities in utility SEM programs or opportunities for them to save energy through distribution system efficiency improvements.
[7] Depending on the SEM program and evaluation objectives, a facility's energy use may include consumption of a single fuel or multiple fuels. Evaluation of savings for multiple fuels is discussed in Section 4.
[8] It may be possible to use facility interval consumption data to estimate energy and peak demand savings Evaluators should consult the peak demand and time-differentiated energy savings protocol (Stern 2013) for guidance about estimating peak demand savings.

- *Facility-level data on energy consumption, production output,[9] and weather[10] for industrial facilities or on energy consumption, weather, floor area, and occupancy or utilization for large commercial buildings are available for the baseline and reporting periods*. Analysis of facility energy consumption, as opposed to analysis of end-use consumption, is recommended for several reasons. First, SEM often affects multiple energy end uses, so only by analyzing whole-facility energy consumption data can evaluators be sure to measure all SEM savings. Second, even if all affected energy end uses could be identified, individual metering may be prohibitively costly. Third, there may be interactive effects between SEM activities that are not recognized or are difficult to measure. Facility energy consumption will capture all of the interactive effects. In addition to facility energy consumption, data on the principal drivers of facility energy consumption, such as output and weather, must also be available for the baseline and SEM reporting periods to perform the savings analysis.
- *Evaluators have sufficient understanding of energy consumption at the facility to construct a valid facility energy consumption model*. Evaluators must also understand the relationships between facility energy consumption and the principal drivers of energy consumption to develop valid energy consumption models. An incomplete understanding increases the risk of incorrectly specifying the baseline regression model. Often, information about facility energy consumption and SEM program activities can be obtained through SEM project completion reports or through interviews with facility energy managers or SEM program implementation staff.
- *Expected energy savings are sufficiently large to be detected with a statistical analysis of the available data*.[11] Evaluators should only apply this protocol when there is an acceptable likelihood of detecting savings using statistical analysis. SEM programs may save substantial amounts of energy, but the savings may only be a small percentage of the facility's consumption and may be difficult to detect statistically. Evaluators can perform a statistical power analysis using baseline energy consumption data to estimate the probability of detecting the expected savings (also known as the study's statistical power).[12]

---

[9] Production is a good or output that the facility produces, measured in physical units (e.g., gallons, meters) per time period. Examples of production include gallons of water treated at a water sanitation facility, hundreds of board feet at a lumber mill, and pounds of carrots at a food processing facility. A good or output may be final or intermediate. An intermediate good becomes an input in another production process at the facility. A final good does not undergo additional processing at the facility. Sometimes only intermediate output data may be available for evaluation.
[10] Data on local weather conditions, including outside air temperature and humidity at appropriate time intervals, should be collected.
[11] SEM programs have saved between 1% and 8% of energy consumption; many had savings goals of about 5%. The range of realized savings represents savings as a percent of consumption for all participating facilities, but often individual facilities saved more than 8%. See CEE (2014b), DNV (2014), Energy 350 (2014), Cadmus Group (2013), and Navigant Consulting (2013). By "sufficiently large," it is meant that savings are large enough to detect, given the number of observations, the variability of energy use, the correlation of energy use, and the availability of information to explain the variation in energy use. Most social scientific studies and program evaluations are designed to achieve statistical power—the probability of detecting a true program effect—of at least 80%. See List et al. (2010). Section 3 discusses the concept of statistical power and application to SEM program evaluations.
[12] ASHRAE (2014) recommends conducting a fractional savings uncertainty analysis, which is similar in concept to a statistical power analysis.

When one or more of the above conditions is not satisfied, other analytic approaches involving building simulations, engineering spreadsheet models, or collection and statistical analysis of consumption data for selected individual facility processes may be appropriate. Such approaches fall outside the scope of this protocol, and readers are encouraged to consult the International Performance Measurement and Verification Protocol (IPMVP) and measure-specific measure level UMP evaluation protocols for further guidance.

2.2 Relationship to Existing and Forthcoming Evaluation Protocols

Two existing evaluation, measurement, and verification (EM&V) protocols address estimation of energy savings from utility SEM programs in large commercial and industrial facilities. A third will be released in 2017 by the DOE.

The first protocol is Option C of the EVO (2012), which applies to comprehensive energy management programs affecting multiple energy-using systems in a commercial or industrial facility. Option C describes analysis of metered energy consumption at the whole-facility or sub-facility levels. Specifically, the IPMVP recommends:

- Applying Option C when the expected energy savings are large relative to the unexplained variation(s) in energy consumption[13]
- Conducting periodic site visits to the facility to identify changes in static factors that may require adjustments to baseline energy consumption
- Estimating baseline energy consumption using regression of baseline period energy consumption as a function of outdoor dry-bulb temperature, production, or occupancy
- Using 12, 24, or 36 months of continuous energy consumption data to estimate the baseline regression model.

The second protocol is the *Superior Energy Performance Measurement and Verification Protocol for Industry* (SEP M&V) (DOE 2017b), which defines procedures for determining compliance with the energy performance requirements of DOE's SEP Program.[14] The SEP M&V Protocol prescribes the following for verifying that a facility meets the requirements for SEP certification:

---

[13] IPMVP recommends applying Option C when savings are expected to be 10% or more of consumption. IPMVP's recommendation is a rule-of-thumb and does not consider the number or frequency of baseline period observations or the amount of unexplained variance of facility consumption.

[14] Utility-administered SEM programs and the DOE SEP Program differ in several ways. First, SEP is a certification program; thus, participants must demonstrate compliance with specific program requirements to be certified. While both programs seek to achieve lasting reductions in energy consumption or energy consumption intensity, SEP requires implementation of a specific energy management system that meets ISO 50001 standards. Most utility- or program-administered SEM programs do not have specific energy management system requirements. Second, SEP covers facility consumption of all energy, while most SEM programs focus on one (e.g., electricity) or sometimes two (e.g., electricity and natural gas) energy types. Third, to qualify for certification under SEP, a facility must satisfy specific criteria on the accuracy of savings estimates. As a consequence, the SEP protocol is more prescriptive about methods for estimating and validating savings than this protocol.

- Conducting top-down analysis of facility energy consumption, as opposed to analysis of specific energy end uses
- Defining facility boundaries that do not change between the baseline and reporting periods
- Defining baseline and reporting periods of at least 12 consecutive months each
- Accounting for all types of energy consumed within the facility boundaries, unless the energy type accounts for 5% or less of total primary energy consumption (in which case it may be justifiable to be ignored)
- Using only data in the estimation that can be independently verified and obtained from precise control and/or measurement systems
- Using statistical models to determine baseline or normalized energy consumption
- Estimating the SEP Energy Performance Indicator, which indicates the percent energy performance improvement
- Conducting a bottom-up analysis and comparison to assess the plausibility of top-down energy savings and performance improvements.

The third protocol is the 50001 Ready Protocol (DOE 2017a), which will be released by the DOE in 2017. Based on the SEP M&V protocol, the 50001 Ready Protocol will allow for determination of energy savings (and carbon emissions reductions) for single or multiple energy types consumed by a facility; however, when used within an ISO 50001-compliant energy management system, the savings determination must include all energy types. The 50001 Ready Protocolwill provide guidance for quantification of energy performance improvement as facilities attain DOE's recognition for being conformant to ISO 50001. Additionally, the 50001 Ready Protocol can serve as a platform on which state and regional SEM program administrators and regulators can build for the specific context of their energy savings and emissions reductions programs.

In general, this UMP evaluation protocol recommends the use of procedures similar to those in the IPMVP option, but provides greater guidance on how to address the specific challenge of determining and evaluating energy savings achieved through SEM.

3 Savings Calculations

This section provides a brief overview of the recommended approach for estimating SEM program energy savings and then describes the step-by-step process for estimating savings.[15]

3.1 Overview of SEM Facility Savings Estimation

Facility energy savings or changes in energy consumption intensity from SEM should be estimated by comparing the facility's metered energy consumption (or energy consumption intensity) during the reporting period with the facility's adjusted baseline during the same period—what its energy consumption (or energy consumption intensity) would have been had SEM not been implemented. The adjusted baseline is a counterfactual, and it must be estimated using baseline period data.

Figure 1 illustrates the estimation of SEM energy savings, showing both metered energy consumption and the adjusted baseline. Savings are shown as the cross-hatched area between the adjusted baseline and metered energy consumption. For simplicity, this example does not differentiate among SEM capital projects, operations, maintenance, and behavioral measures.

[15] Many programs have sought additional savings opportunities from an ISO 50001-conformant EnMS, and so programs may seek to include EnMS as a program element or a potential second category of SEM program. Facilities and companies that have obtained or are seeking ISO 50001 conformance or certification should use the 50001 Ready Protocol (alternatively, the SEP M&V protocol) to determine energy savings. The SEP program provides requirements regarding the determination and verification of energy performance improvement for its ISO 50001-based certification program through the SEP M&V Protocol (DOE 2017b) and SEP Certification Protocol (DOE 2016b).

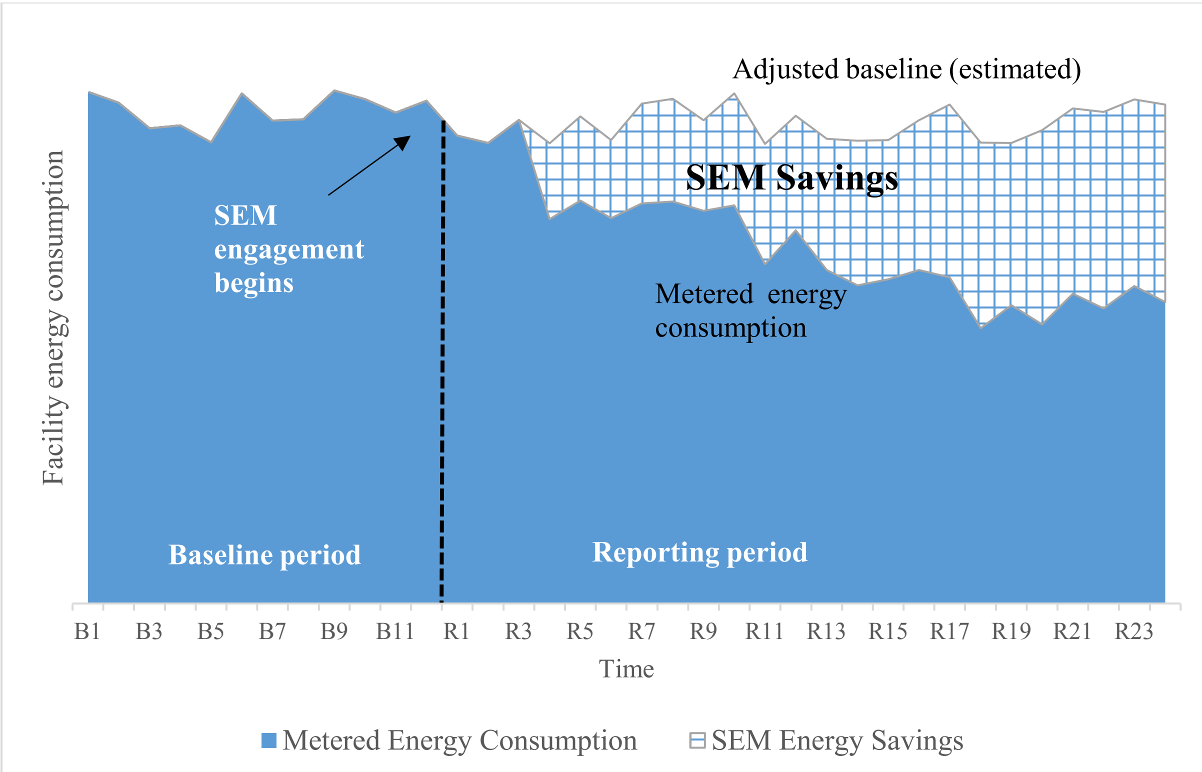


Figure 1. Estimation of SEM energy savings

Notes: Figure 1 illustrates some expected savings trends for an SEM program facility. During the first few periods of the reporting period, the facility may save little or no energy as the facility plans and begins to implement SEM. Then the facility begins to save energy, followed by a period of plateauing savings. As SEM program facilities are expected to continue to implement efficiency measures, savings begin to increase again around period 10.

The adjusted baseline should be estimated using facility energy consumption data from the baseline period, which should not reflect the SEM program impacts the evaluator wishes to measure. Typically, the baseline period precedes the facility's SEM implementation.

Using regression, the evaluator should adjust the baseline energy consumption for differences between the baseline and reporting periods in output, weather, occupancy, or other measured variables affecting the facility's energy consumption. Section 4 of this protocol describes five specific regression methods for estimating the adjusted baseline and savings.

This approach for evaluating facility savings from SEM programs will yield accurate savings estimates if the following conditions are met:

- No omitted variable bias (no confounding variables): The regression does not omit any key variables affecting energy consumption. Specifically, the model controls for all variables that affected energy consumption and that were correlated with SEM implementation.
- No significant measurement error: The model's independent variables were not measured with minimal error.

For example, omitted variables could bias the SEM-savings estimates if an industrial facility experiences a degradation in the quality of production inputs during SEM, causing energy consumption per unit of output to increase, and the change in input quality is not accounted for. The change in input quality would be a confounding factor, causing downward bias in the estimated savings.

The evaluator should take steps to minimize the potential for omitted variables and measurement error. These include collecting data on the principal factors affecting facility energy consumption and conducting statistical tests addressing whether the conditions required for unbiased estimates hold. However, temperature and other candidate predictor variables may only be known with error, in which case an error-in-variables estimation approach such as instrumental variables two-stage least squares should be considered.

SEM may involve implementation of OM&B measures and capital projects, and evaluators may wish to isolate savings from OM&B measures. This protocol discusses estimation of these savings below.

For some facilities, it may be necessary for the evaluator to make ad hoc adjustments to the baseline to capture impacts on energy consumption that cannot be modeled statistically. These are referred to as "non-routine" adjustments (IPMVP 2012). Section 4 of this protocol discusses the use of non-routine adjustments.

To estimate SEM program energy savings, evaluators should follow these steps:

1. Develop research design (includes sample design, if applicable)
2. Collect documentation and prepare required data
3. Define baseline and reporting periods
4. Specify regression model
5. Estimate regression model
6. Estimate and document savings
7. Report results.

To make the evaluation successful, evaluators should work closely with program administrators and implementers, especially with regard to research design and data collection. Ideally, evaluators should coordinate with program administrators and implementers during the program design phase to ensure that data required for evaluation will be collected. However, as the early involvement of evaluators will not always be possible, program administrators should familiarize themselves with the guidelines about research design and data collection to make sure their programs are evaluable.

The remainder of this section discusses each of these steps.

3.2 Develop Research Design

Research design involves developing the approach for selecting the analysis sample, collecting data, and estimating the savings. Evaluators should carefully design the evaluation, ideally working closely with program managers and implementers, to ensure that the evaluation objectives can be achieved. Involving evaluators early will increase the likelihood that the evaluation will achieve its objectives and obtain accurate savings estimates.

During the research design process, evaluators should determine the following:

- *Evaluation goals*. Evaluators and program managers should agree on goals for the evaluation to ensure that the required data can be collected and that the evaluation answers the program administrator's research questions.
- *Variables necessary to model facility energy consumption, so the means to collect the required data can be put in place*. For industrial SEM programs, verifying the availability of data is an important step as some industrial utility customers may not have the data in an accessible format or may not be willing to share data on facility inputs or outputs. For commercial buildings, verifying the availability of occupancy data and the frequency of available data represents necessary steps, as occupancy can be an important explanatory variable.
- *Required sample sizes in terms of facilities and amount of data for each facility*. The sample size calculation will depend on the program design, evaluation objectives, and frequency of available energy consumption data. Specifically, the sample size calculation will differ for the following levels of disaggregation:
    - A regression of energy consumption involving a single facility. The evaluator should determine the number of baseline period observations and the number of reporting period observations of energy consumption required to detect the expected facility savings.
    - A regression of energy consumption for a census of multiple facilities that participated in an SEM program. In this case, the evaluator should determine both the number of observations and the number of facilities that must be sampled, accounting for within-facility correlation of energy consumption.
    - Individual regressions of energy consumption for multiple facilities from a sample of the population. In determining the number of facilities to sample, the evaluator should account for error from both sampling and modeling.
- *The likelihood of detecting savings at the desired levels of statistical confidence and precision for evaluations that will be performing facility-level analysis*. If there is a low probability of detecting savings using statistical analysis of facility consumption, the evaluator should consider other approaches for estimating savings, such as statistical analysis of sub-meter data.
- *Expectations for changes in the facility production process or input characteristics that would substantially alter facility energy consumption*. It may be necessary for evaluators to collect data on these changes to obtain an accurate estimate of savings.

3.2.1 Define the Facility and Energy Consumption Boundaries

As part of the research design, the evaluator also should define the energy consumption boundaries of each facility. As noted above, the facility is the unit of analysis and the area over which energy consumption will be measured and analyzed. A facility could be an entire industrial or large commercial site or a subset of a site. For example, an industrial site may comprise several industrial processes located in different buildings that are separately metered. In this case, a facility could be defined as the entire site or one or more buildings onsite.

Evaluators should attempt to define the facility boundary so that the boundary covers all of the SEM energy savings. However, in some cases, evaluators may choose to define the facility boundary more narrowly—only including a subset of energy uses affected by SEM activities—or more broadly—including energy consumption of some activities or facility areas unaffected by energy consumption—to obtain valid savings estimates. The choice of facility boundary may involve tradeoffs and depend on considerations of not just the facility areas affected by SEM activities, but also on the availability of energy consumption and other facility data such as facility production, the evaluator's ability to detect the savings using statistical methods, and evaluation objective. For example, an evaluator may face a tradeoff between obtaining a comprehensive facility savings estimate and a precise savings estimate. By defining the facility boundaries broadly, the evaluator's analysis may result in an estimate of savings for all SEM implementation activities but because of noise in the data, the estimate may be imprecise. Alternatively, by defining the facility boundary narrowly, the evaluator's analysis may exclude the savings of some implementation activities but reduce noise in the data and achieve a more precise estimate of savings implemented in that narrower boundary.

However the facility is defined, the evaluator should define the facility boundaries consistently, and should collect measurements of facility energy consumption and other key variables consistently over the study. In addition, if the facility is defined as a subset of a site, the subset should not have significant interactive effects with other parts of the site, and the subset should have separately metered consumption for all energy types evaluated.

3.2.2 Identify On-Site Energy Uses

As a facility may consume multiple types of fuels, the evaluator should identify the facility's consumption of different energy types or fuels (e.g., electricity, natural gas, fuel oil) and the types of energy consumption expected to be affected by SEM.

Also, a facility may consume some fuels delivered from outside suppliers and others generated onsite. For example, many large commercial buildings rely exclusively on utility-supplied electricity for their power needs. But some large commercial buildings also generate some power onsite using renewable generation or combined heat and power technologies. The same holds true for many industrial facilities, which may rely on a combination of delivered and onsite generation of electricity. The evaluator must understand and account for the facility's energy sources to ensure that the measurement of facility energy consumption is accurate.

More formally, in a given time period, consumption of energy will be the sum of delivered and onsite production of energy minus any exports and changes in onsite inventory of the energy:

$$\text{Energy consumption} = \text{Onsite Generation} + \text{Deliveries} - \text{Exports} - \text{Inventory Changes}$$

Some evaluators may find it helpful to draw a system diagram showing the flow of energy through the facility. A well-done system dynamics "stock and flow diagram" can make clear what is happening with energy and what is being assessed.

Some factors may not be relevant for certain types of energy (for example, inventories for electricity unless the facility has electricity storage capabilities). As the equation shows, however, when one or more of onsite generation, exports, and storage of energy are feasible, data on all relevant elements (not just delivered energy) are required. Also, deliveries of energy could fall, but consumption could increase if onsite generation increased or if exports decreased by a greater amount. Focusing on just electricity delivered by the utility might produce misleading results.

At the outset, the evaluator also should determine the energy types for which savings will be measured and whether savings from multiple energy types should be combined to determine overall savings. The evaluator should be aware of a facility's potential to substitute between different types of fuels. Substitution of, for example, natural gas for electricity—for some energy end uses—may result in a reduction in facility electricity consumption, but, depending on the SEM program objectives, this reduction may not qualify as energy savings. Moreover, fuel substitution may not result in a reduction in overall site energy consumption.

When a facility can substitute between fuels, evaluators should conduct individual consumption analyses for the substitutable fuels or convert consumption of the substitutable energy types to a common energy unit, such as joules, kWh, or British thermal units (Btu), and analyze the combined consumption. This conversion is necessary for a facility that can switch between electricity and natural gas, which might mean that some electric savings are offset by increases in gas, which would not be detected by a single-fuel electricity model.

Finally, evaluators should determine whether total savings should be calculated in terms of delivered energy or primary energy, which accounts for any energy consumed in the production and transport of delivered energy.[16]

3.2.3 Conduct Statistical Power Analysis

During development of the research design, evaluators should conduct a statistical power analysis to determine the study's likelihood of detecting the expected savings. The probability of detecting savings is known as the *statistical power* of the study and is a function of the following:

- The expected SEM savings as a percent of consumption;
- The variability of facility energy, as measured by the coefficient of variation (CV)[17] of facility energy consumption;
- The probability of concluding savings occur when there are none (also known as the probability of making a type I error and the statistical significance level);[18]

---

[16] For guidance about the calculation of primary energy, see Deru and Torcellini (2007) and Annex B of DOE (2017b).

[17] The CV of a random variable is the ratio of the sample standard deviation to the sample mean.

[18] A Type I error occurs when a researcher rejects a null hypothesis that is true. Statistical confidence equals 1 minus the probability of a Type I error. A Type II error occurs when a researcher accepts a null hypothesis that is false. Many researchers agree that the probability of a 5% Type I error and a 20% Type II error is acceptable. See List (2010).

- The number of energy consumption observations for the baseline period;
- The number of energy consumption observations for the reporting period; and,
- The correlation of facility energy consumption over time

A study may have low statistical power because the expected savings are small, there is substantial unexplained variability in the facility's energy consumption, or the number of observations in the baseline or reporting period are small. Evaluators also can use a statistical power analysis to determine the number of baseline and reporting period observations necessary to achieve a desired statistical power.

Statistical power can be calculated in two ways. First, evaluators can calculate it analytically, using standard formulas that require as inputs the bulleted items above.[19] The statistical power formula will vary, depending on the study's design. Evaluators who conduct analysis of individual facilities will need to input the number of energy consumption measurements in the baseline and reporting periods as well as facility energy consumption characteristics. Evaluators who conduct a panel regression analysis will need to input the number of energy consumption measurements in the baseline and the reporting periods, energy consumption characteristics, and the number of facilities in the analysis sample.

Second, evaluators can assess statistical power numerically, using simulations. This approach will work well if evaluators have high frequency consumption data (maximum intervals of a week) for at least one year of the baseline period. Evaluators should simulate the expected program savings for a portion of the baseline period, say, the second half, by adjusting the data accordingly. Then, for the remainder of the baseline period (e.g., the first half), evaluators should sample observations randomly with replacement, estimate a baseline consumption model with the sampled observations, and estimate savings for the simulated reporting period. Then evaluators should repeat this exercise a large number of times, e.g., 200 or more, calculate the distribution of estimated savings, and determine the percentage of iterations that the estimated savings were greater than zero. This percentage equals the statistical power of the study—the probability of detecting the expected savings when the true savings equal the expected savings.

3.3 Collect and Prepare Required Data

This protocol recommends using regression analysis to estimate the adjusted baseline because regression can account for changes in factors affecting facility energy consumption between the baseline and reporting periods. For example, the adjusted baseline should account for increases in output or space conditioning demand during the SEM reporting period relative to the baseline period. It is therefore essential that evaluators collect data on the principal time-varying drivers of facility energy consumption. Specifically, evaluators should collect the following data to estimate SEM program savings:

- Facility energy consumption;
- Facility production outputs for industrial facilities;

[19] See Frison (1992) or List (2010) for specific power calculation formulas. Evaluators can conduct statistical power calculations using SAS, Stata, and R software.

- Facility occupancy for commercial buildings;
- Local weather;
- Facility shutdowns or closures;
- SEM measures and implementation schedules;
- Other efficiency measures; and
- Changes in facility or building operations or production unrelated to SEM, but affecting energy consumption.

For some facilities, it may be necessary to use proxies when occupancy data are unavailable. For example, with respect to primary and secondary schools, it is unlikely that data on building occupancy will be available; however, evaluators can use the calendar of school openings and closings to model whether a school building was occupied during a particular day.

Also, evaluators should be aware of any significant one-time changes in the facility unrelated to SEM implementation. Evaluators should collect data on these non-routine changes and determine how best to account for their effects on facility energy consumption. For example, a facility may have experienced a change in the quality of production inputs that necessitated an adjustment to the reporting period consumption data.

3.3.1 Energy Consumption Data

Evaluators should collect data on energy consumption during the SEM baseline and reporting periods for all of the energy types the SEM program will evaluate. The evaluator should collect these data from the utility supplier or the program administrator.

Evaluators should attempt to collect daily facility energy consumption data for analysis. If available, hourly energy consumption data can be aggregated to the daily level. Collecting high-frequency data is encouraged for several reasons:

- High-frequency data usually increase the probability of detecting energy savings. For example, a recent study for the Bonneville Power Administration (BPA) found a strong positive correlation between the frequency of a facility's energy consumption data and the statistical significance of SEM energy savings at the site.[20]
- High-frequency data may provide greater insights about SEM program effects. For example, with daily energy consumption data, it may be possible to identify the effects of SEM measures intended to save space conditioning energy consumption by correlating daily energy consumption with daily cooling degrees.[21] In addition, by using daily energy consumption data, it may be possible to identify the specific effects of measures designed to impact weekday (production) or weekend (non-production) operating modes.

[20] See Cadmus Group (2013).

[21] The evaluator should also consider the costs of collecting high-frequency data, as collecting these may not be cost-effective. Further, just because high-frequency data increase the probability of finding significant savings, the point estimate of savings may not differ. An alternative to collecting high-frequency data would be to increase the number of sites to improve the overall program-level estimate.

- It may be possible to observe a wider variety of facility operating conditions with high-frequency data, which may mitigate some of the limitations from estimating savings based on shorter baseline or reporting periods.

Often, a binding constraint on an evaluator's ability to analyze high-frequency energy consumption data is the unavailability of other analysis data at the same or higher frequencies. For instance, an SEM-participating facility may be unable—or unwilling—to provide sensitive, high-frequency occupancy or production data. Also, some kinds of data—including production from "batch processes" that occur over multiple days or energy consumption for some fuels (e.g., gas, propane, coal)—often are unavailable at daily frequencies. In addition, there may be a delay before the facility collects such data and provides it to the evaluator. When energy consumption is reported at a higher frequency (e.g., daily) than other analysis variables (e.g., monthly), it may be necessary to aggregate energy consumption and other data to the minimum frequency of the secondary analysis variables.

Another possible situation is that energy consumption data are reported at different frequencies during the baseline and reporting periods. If baseline period data are reported at a higher frequency, the evaluator may use the high-frequency data to estimate the adjusted baseline, aggregating the estimates of adjusted baseline energy consumption to the reporting-period data frequency to calculate savings. It is more likely, however, that baseline-period energy consumption will be reported at a lower frequency than reporting-period energy consumption due to recent advances in high-frequency metering deployment. In this case, the adjusted baseline has a monthly frequency and it is necessary to aggregate the reporting period data to the baseline data's frequency to estimate savings. Another potential solution to this problem involves establishing a new baseline period that only includes consumption reported at the higher frequency.

3.3.2 Variables Affecting Facility Energy Consumption

Evaluators should collect data on the principal drivers of facility energy consumption. In industrial facilities, the principal energy consumption drivers typically will be production outputs and weather. In commercial buildings, the principal drivers most likely will be occupancy and weather. In commercial buildings such as offices, space conditioning usually is the single largest energy end use, accounting for over 40% of total building consumption.[22] While industrial processes that are not sensitive to weather often account for the large majority of energy consumption at industrial facilities, weather-sensitive energy consumption for space conditioning or industrial refrigeration or heating can still be significant, and evaluators should collect weather data to account for these end uses.

Accuracy of the savings estimates may be improved if evaluators collect data on building closures for commercial buildings and on full- or partial shut-downs for industrial facilities. For example, incorporating information about school holidays and occupancy into energy consumption models can significantly improve the model's accuracy. Similarly, an industrial facility will likely have very different energy consumption when it is idle than when it is open

[22] Energy Information Administration (2008).

but producing a low volume of output. Knowledge about industrial facility operating conditions can be used to improve the accuracy of the energy savings estimates.

3.3.3 SEM Program-Related Facility Activities

At a minimum, evaluators need to collect sufficient information about the program's implementation to define the baseline and reporting periods, and to estimate the adjusted baseline.

Evaluators also should collect the following data on implementation of SEM program-related activities at a facility:

- Company background;
- Facility background, including location, building type, outputs for industrial sites, occupants for commercial buildings, and any changes in facility operations;
- Descriptions of key drivers of energy consumption;
- Results of any facility energy efficiency opportunity assessments or audits;
- SEM program implementation start and end dates, and the expected energy savings;
- Description of SEM facility boundaries, program design, objectives, and milestones;
- Description of the facility-level SEM framework, including implementation details of relevant SEM elements (e.g., energy policy, type and scope of trainings, and process for measuring energy performance improvement);
- Descriptions of SEM energy efficiency measures and activities;
- Descriptions of other energy efficiency capital and retrofit projects, including detailed M&V documentation implemented during the baseline or reporting period; [23]
- Descriptions of any changes in facility or building operations and maintenance, unrelated to the SEM program during the baseline and reporting periods; and
- Descriptions of SEM and capital project energy savings estimations, and assumptions used in those estimations.

Many program administrators or implementers present this facility information in an annual SEM program report or in a register of implemented projects. Evaluators should use these data to build valid models of facility energy consumption and to assess whether the evaluation savings estimate is reasonable, given the actions taken at the facility. Also, evaluators should use information about how the utility SEM program was implemented at the facility to put the savings estimates into context, specifically when assessing the program's success in encouraging organizational and operational changes to improve the facility's energy management and efficiency.

[23] Description should include prior implementation of any SEM, capital, and retrofit projects during the previous five years.

3.3.4 Facility Energy Manager or SEM Implementer Interviews

After reviewing SEM documentation, the evaluator may have outstanding questions about the facility's operations, energy consumption, or SEM activities. For example, the evaluator may be unclear about the implementation date of a particular SEM activity or a change in facility operations. The evaluator also may need additional information to develop a valid model of facility energy consumption or to make non-routine adjustments. In such cases, this protocol strongly recommends evaluators request clarification from a facility energy manager or from SEM implementation staff.

Additionally, evaluators may wish to conduct interviews with energy managers or implementation staff for some or all evaluated facilities. Interviews, which may be necessary for a process evaluation, allow the evaluator to make significant improvements to the facility energy consumption models.

Evaluators should tailor interviews with facility energy managers or program implementers to reflect a particular facility and SEM program. The following list of generic, SEM-related interview questions can be modified to fit an evaluator's specific needs. The first two questions can help assess the program participant's SEM awareness and engagement before participation, and provide important context for measuring program impacts:

- What is your current understanding of SEM? Before participating in the SEM program, was your facility aware of SEM? If so, please describe your previous awareness and understanding of SEM.
- Which, if any, of the 13 CEE minimum SEM elements did your facility implement before participating in the SEM program?[24]
- Can you confirm that the following SEM program activities were implemented? Are they still in place?
- What kind of energy was the SEM program intended to save? How much energy did you expect to save? How much energy did you expect to save as a percent of consumption? Which SEM activities directly produced energy savings?
- Since participating in the SEM program, have there been any substantial changes to the facility (e.g., changes in floor area, new production lines)? If so, please describe.
- Since participating in the SEM program, have there been any changes in operating hours/schedules? If so, please describe the operating hours/schedules before and after participating in the SEM program.
- Since participating in the SEM program, has there been any change in facility management or staffing? If so, please describe those changes and how they impacted the operation of the facility before and after participating in SEM.

[24] Evaluators should keep in mind that most program participants will be unfamiliar with the CEE minimum elements and should be able to ask about implementation of the minimum elements without referencing them by name.

- Since participating in the SEM program, have there been any replacements or installations of new machinery or equipment? If so, please describe the changes.
- Have there been any significant changes in production levels since implementing SEM?
    - How did these changes affect energy consumption?
    - What was the reason for these production changes (e.g., does production vary seasonally)? Are the production changes permanent? If not, when do you expect them to change again and to what level?
    - Did the program have any role in this change? If so, what was its role? Are these changes permanent?
- Since participating in the SEM program, have you changed the product line or added any different products to your production line? If so, did the program have any role in how you set up production of these new products?

3.4 Define Baseline and Reporting Periods

The baseline period should be sufficiently long to cover the range of operating conditions that the facility experienced prior to SEM implementation and to provide enough data to precisely estimate the coefficients of the energy consumption regression. This protocol recommends collection of a full year of baseline data. A full year is usually sufficient to capture any changes in energy consumption related to weather, seasonal market demand for facility output, and facility closures and schedules.

In some cases, a baseline period of a year may be unfeasible. In these situations, it may be possible to use the shortened baseline period if it is representative of conditions during the reporting period. For example, it may be possible to use a baseline of a few months to estimate savings for an industrial facility without weather-sensitive energy consumption and that produced output levels within the same range during the reporting period. In contrast, a baseline of a few months would be insufficient for a large office building with very weather-sensitive energy usage. Such facilities require a baseline period that includes summer, winter, and shoulder months.

The baseline period and reporting period also should exhibit similar ranges of facility operating conditions. It is unnecessary for the operating conditions to overlap 100%; however, the evaluator should be confident that the regression model will predict energy consumption accurately over the range of reporting period conditions.

If the baseline period and reporting period do not exhibit similar ranges of conditions, the energy consumption regression model estimated with baseline period data may not accurately predict the adjusted baseline. For example, if a food processing facility produced different outputs during the baseline and reporting periods (e.g., frozen vegetables during the baseline period and frozen fruits during the reporting period), and these outputs required different amount of energy per unit of output, accurately estimating the adjusted baseline would be difficult. Similarly, an evaluator will be unable to accurately estimate the adjusted baseline for a large office building during peak-cooling summer months if the baseline period does not include days with similar temperature ranges.

This protocol recommends evaluators follow the guidelines in Section 6.4.2 of the *SEP M&V Protocol* when establishing the similarity of baseline and reporting period conditions. According to the SEP M&V protocol, the means of the adjustment model's variables during the reporting period "should fall within both:

- The range of the baseline period data used to estimate the model.
- Three standard deviations from the means of the adjustment model variables during the baseline period.

Any outliers excluded when estimating the baseline consumption model should also be excluded when calculating the valid quantitative range of the model-related variables."[25]

3.4.1 Redefining the Facility Baseline

An important issue for programs running for longer than one year concerns the validity of the original baseline. This protocol recommends that evaluators maintain the original facility baseline as long as the baseline remains valid. Specifically, evaluators should continue to use the original baseline if the baseline and reporting periods have similar operating conditions, not counting SEM program effects.

During the reporting period, however, some facilities may experience significant changes in operations—unrelated to SEM—that affect energy consumption. These changes may invalidate the original baseline and necessitate selecting a new one. Some SEM program administrators and implementers have reported redefining baselines for many facilities after two or more years of SEM engagement because the original baselines were no longer valid due to changes in operations, occupancy, and product mix. However, even if facility operations remain unchanged, evaluators may want to establish a new baseline to take advantage of new data that has become available as the new data may make it possible to build a more accurate baseline model.

In these cases, this protocol recommends evaluators consider selecting a new baseline period with operating conditions similar to those of the reporting period. Also, it may be necessary to select a baseline period that includes some SEM program activity. For example, if a facility made significant changes to its production process or started producing new kinds of output after the start of SEM implementation, the evaluator would be unable to use the period preceding SEM implementation as a baseline. Instead, the evaluator could use the 12 months immediately following the change in facility operations as a baseline for measuring energy savings during subsequent program years.

When the evaluator redefines the baseline and the new baseline includes SEM activity, the evaluator will measure SEM program effects relative to the more efficient baseline. The savings estimate will exclude the effects of any measures implemented before or during the redefined baseline period. Only incremental SEM savings—savings from measures implemented since the end of the new baseline period—will be measured.

[25] DOE (2017b), p. 23.

3.5 Specify Energy Consumption Regression Model

Next, the evaluator will need to specify the regression model for the facility's energy consumption. This involves defining the dependent variable, determining which independent variables will be included in the model, and determining each independent variable's functional relationship to the dependent variable. The evaluator also will need to specify the assumptions about the properties of the model error term and test those assumptions.

To be valid, a regression model need not exactly represent the physical energy consumption relationships in the facility. At most SEM facilities, particularly industrial facilities, these relationships are likely too complex to be represented exactly. The frequency of available data also may not allow for the estimation of such a model, even if it could be developed.

Instead, a valid regression model accurately predicts the facility's adjusted baseline and yields an accurate estimate of facility energy savings. Evaluators can use statistical methods in constructing the regression model. These methods can help the evaluator identify relationships in energy consumption data not evident through engineering analysis. This does not mean evaluators should ignore knowledge of facility energy consumption relationships; rather, understanding the facility's end use will likely increase the energy consumption model's validity.

As a first step to developing an energy consumption regression model for a facility, this protocol recommends evaluators carefully review documentation about the facility's energy consumption. In addition, evaluators should review the specification and estimation results of the implementer's energy consumption model. These reviews should inform construction of the evaluator's model and, in fact, the implementer's model may serve as a starting point for constructing the evaluation model.

3.5.1 Selecting the Dependent Variable

The model-dependent variable either will be facility energy consumption per unit of time (e.g., day, week, month) or facility energy consumption intensity per unit of time. In industrial facilities, energy consumption intensity is usually defined in relation to output, whereas energy consumption intensity in large commercial buildings is usually defined in relation to floor area.

The choice to use energy consumption or energy consumption intensity as the dependent variable will depend on the evaluation's primary objective (i.e., to measure energy savings or reductions in energy consumption intensity). Section 4 of this protocol discusses the estimation of energy consumption and energy consumption intensity regressions. It is possible, however, to obtain estimates of energy savings using either specification.

3.5.2 Selecting Independent Variables

The energy consumption regression model specification should be determined on the basis of engineering knowledge about the facility's energy consumption and statistical diagnostics and testing.[26]

Information about physical energy consumption relationships at a facility usually can be obtained through a facility project completion report or through interviews with plant managers or program implementers. Engineering knowledge about energy as an input to the production process may tell the evaluator that the energy consumption has a specific relationship (e.g., linear or nonlinear) with output.

For example, production may require less energy per unit of output as the production level increases. In this case, the evaluator should select a functional relationship for energy consumption with output that reflects this nonlinear relationship. Similarly, at a water treatment and sanitation facility, groundwater may be pumped from different depths, and some pumps may use more energy per gallon of water pumped than others. The estimating relationship should reflect these differences, especially if the volume of water pumped from different depths varies over time.

Plotting facility energy consumption against time and each of the candidate independent variables provides a good starting point. These plots can identify variables that have strong relationships with energy consumption, as well as the nature of those relationships. The plots also may suggest which candidate variables are highly correlated and collinear. Multiple variables, however, may exhibit similar relationships with energy consumption; therefore, more sophisticated methods for selecting variables may be required.

Evaluators can use statistical methods to select independent variables, which can help the evaluator identify variables correlated with energy consumption that engineering analyses did not identify. Statistical methods also can be used to determine whether higher-order terms (i.e., squares and cubes) or interactions between independent variables should be included as regressors.

There are well-developed, automated statistical procedures of varying sophistication for selecting model-independent variables. These methods typically involve estimating a large number of regression models that include different variables or assume different model parameter values from the feasible parameter space, and selecting the variables and parameters that produce the best regression fit.

For example, evaluators can use statistical methods to determine the appropriate change-point temperature for modeling a facility's space heating or space cooling energy consumption. Evaluators can find the heating degree and cooling degree base temperatures that best explain a commercial building's energy consumption by running regression models with different heating

[26] DOE has a regression-based tool for helping researchers in assessing a facility's energy performance and identifying the variables affecting a facility's consumption. The tool is available online: https://ecenter.ee.doe.gov/EM/tools/Pages/EnPI.aspx.

degree and cooling degree base temperatures, and then selecting the base temperatures that yield the best model fit.[27]

As another example, evaluators can use forward-selection, backward-selection, or stepwise-selection regression methods to select model-independent variables. Each method is an automated, iterative process that identifies variables correlated with facility energy consumption. In all cases, the evaluator first identifies candidate variables for the model and a statistical significance level that selected variables must satisfy. The evaluator should select candidate variables based on knowledge about the facility's energy consumption. For most commercial buildings, candidate variables will only include cooling degree days (CDDs), heating degree days (HDDs), and possibly occupancy. The routines differ in whether variables iteratively are added to or removed from the model, and whether added variables can be subsequently removed. Automated variable selection routines can be found in statistical software packages such as R[28], Statistical Analytics Software (SAS), and Stata.

While statistical methods can be useful for choosing model specifications, evaluators should also exercise caution, being careful not to hand over too much control to a computer. One way evaluators can do this is by forcing the model to include certain variables known to influence energy consumption, while testing the appropriateness of including other variables, interactions, or higher-order terms (squares and cubes).[29] Evaluators should consider rejecting model specifications that yield energy consumption relationships that are implausible or counter-intuitive.

Evaluators should try to avoid omitting variables from the model that significantly affect facility energy consumption. Models omitting such variables will be specified incorrectly and the savings estimates may be biased.

3.5.3 Model Error

Specifying the model also requires making assumptions about the properties of the error term. The error term represents influence of unobserved factors on a facility's energy consumption. These assumptions help determine the approach for estimating the model.

Often, evaluators assume the energy consumption regression model satisfies the classical assumptions of an ordinary least squares (OLS) regression model. These assumptions concern:

- The variance of the error term (i.e., the error term has constant variance),
- The independence between the error and the independent variables (i.e., the error term is uncorrelated with model explanatory variables),

[27] Less computationally intensive methods can be used to identify the change point. For example, the evaluator can plot facility energy use against outside temperatures and attempt to visually identify temperature change points. However, if data are noisy or space conditioning accounts for a small share of the facility load, it may be difficult to identify the temperature change points visually.

[28] A software environment for statistical computing and graphics provided by The R Project. https://www.r-project.org/

[29] Chapter 13 of Imbens and Rubin (2015) provides guidance about building valid regression models using automated variable selection procedures.

- Serial correlation of the error term for time series models (i.e., observations are independent over time), and
- Collinearity between the independent variables (i.e., the explanatory variables are not collinear).

Using statistical tests, evaluators should verify that the assumptions hold about the energy consumption regression model error. If the assumptions do not hold, it may be necessary for the evaluator to re-specify the model or to estimate it using a different method.[30] Standard econometric texts describe statistical tests for checking the important assumptions of an OLS model (Greene 2012).

3.6 Fitting the Model

After determining the model specification, the evaluator should select a method for estimating the model. Knowledge about the properties of the model's variables and error should guide the estimation approach. Detailed guidance can be found in most econometrics texts, such as Greene (2012).

3.6.1 Model Fit Tests

After estimating the energy consumption model, the evaluator should assess the model's fit and conduct tests of key model assumptions.[31] Texts by the BPA (2012), the SEP M&V Protocol (DOE 2016c), and standard econometrics texts describe many standard tests of model fit and validation.[32]

When beginning testing, the evaluator should first plot the model residuals, looking for anomalous patterns suggesting omitted variables, auto-correlated errors, or heteroscedastic errors. The evaluator should also inspect the model coefficient of determination (R^2), the regression F statistic, and the signs and statistical significance of the coefficients. The model R^2 indicates the amount of variation in the dependent variable explained by the model-independent variables.

---

[30] For an interesting example of an energy savings analysis of a commercial building that deviates from the standard OLS regression assumptions, see Price (2014)..

[31] Amundson (2013) and Northwest Industrial Strategic Energy Management Collaborative (2013a, 2013b) illustrate several model-specification tests for industrial SEM energy use regressions.

[32] This protocol does not require the baseline consumption model to meet specific values for the model fit tests; however, other protocols have such requirements. As an example, according to Section 6.4.1 of the SEP M&V Protocol (DOE 2017b), a valid model must demonstrate the following:

- An F-test for the overall model fit must have a p-value less than 0.10 (i.e., the overall fit of the adjustment model is statistically significant greater than the 10% significance level).
- All included variables in the model must have a p-value less than 0.20.
- At least one of the variables in the model must have a p-value less than 0.10.
- The R^2 for the regression must be 0.50 or greater.
- The selection of relevant variables in the adjustment model and the subsequently determined relevant variable coefficients are consistent with a logical understanding of the energy use and energy consumption of the facility.

These are reasonable requirements for determining model validity and evaluators may wish to impose all, some, or none of these requirements. If consensus builds in the industry for specific threshold values for these requirements, these values can be incorporated when this protocol is updated.

A low R^2 should be investigated because it indicates that the regression model does not explain much of the variation in energy consumption. Nevertheless, a model with low R^2 may still produce an unbiased, statistically significant savings estimate. The regression F statistic measures the overall statistical significance of the regression and can be used to test whether the model-independent variables have statistically significant effects on energy consumption. The estimated model coefficients should have the expected signs and magnitudes, based on engineering knowledge about the facility. However, the evaluator should keep in mind that a large R^2 or statistical significance is not sufficient to conclude that the model makes valid predictions of energy consumption. The estimated coefficients of an incorrectly specified model may be statistically significant.

To further investigate the model validity, the evaluator also can plot predicted energy consumption against metered energy consumption. The evaluator should verify that the model explains energy consumption at all ranges of output or the weather at which the model is intended to apply.

The evaluator also may be able to test the predictive accuracy of the baseline model by holding out some baseline period observations from the estimation sample. The evaluator can estimate the model with the remaining baseline period observations and then use the model to predict energy consumption for the hold-out observations. A valid model should closely predict the energy consumption during the hold-out intervals.

Finally, the evaluator should check the sensitivity of the regression estimates to changes in any key assumptions. Those assumptions could concern:

- Definition of the baseline and reporting periods;
- Whether variables influence energy consumption and belong in the regression; and
- The functional form of the regression-dependent variable, such as whether the regression specification is linear, logarithmic, or semi-logarithmic.

3.7 Estimating and Documenting Savings

The evaluator should use the estimated regression to estimate the adjusted baseline and then to estimate savings as the difference between the adjusted baseline and metered energy consumption. Section 4 of this protocol describes and illustrates two regression approaches for doing this.

Evaluators should document the method for estimating the energy consumption regression model and energy savings, including the following:

- Period(s) covered by data used to estimate the model;
- Baseline and reporting period definitions;
- Model specification and assumptions;
- Estimation approach;
- Estimates of regression coefficients and standard errors;

- Relevant model fit statistics, including R^2 and F statistics; and
- Calculations used to estimate savings, including any non-routine adjustments to the adjusted baseline.

3.7.1 Estimating Savings Attributable to OM&B Measures

This protocol focuses on estimating overall energy savings from SEM activities, whether from OM&B measures or from capital and retrofit projects. However, as implementation of OM&B measures is an integral component and defining feature of SEM programs, program administrators and regulators may ask for a separate estimate of OM&B savings. Also, other utility programs may claim savings from capital projects, requiring evaluators to obtain a separate estimate of the remaining OM&B savings.

When an SEM program facility only implements OM&B measures, the facility energy savings estimate *is* the estimate of OM&B savings. However, when a facility also implements capital or retrofit measures, evaluators must have an estimate of the capital or retrofit project savings to estimate the OM&B savings.

Evaluators can obtain an estimate of the OM&B savings by subtracting the capital or retrofit project savings estimate from the regression-based facility savings estimate:

$$\text{OM\&B Savings} = \text{Facility Savings} - \text{Capital or Retrofit Measure Savings}$$

The OM&B savings estimate depends on the accuracy of the facility savings estimate and the capital measure savings estimate. The estimated OM&B savings will increase or decrease one-for-one with opposite changes in the estimated capital or retrofit project savings. Thus, any error in the estimate of capital measure savings will result in an opposite and equal error in the OM&B savings. Error in the facility savings estimate also will result in error in the OM&B savings estimate.

Evaluators should be cautious in using this approach to disaggregate SEM savings. First, despite evaluators' best efforts to ensure accuracy, capital project savings may be estimated with significant error. This particularly may be the case for utility programs that rely on deemed savings approaches, as the actual capital project performance may vary greatly from facility to facility. Evaluators may be able to improve the accuracy of the capital project savings estimates through sub-metering of specific facility processes and should consider the expected evaluation benefits and costs of sub-metering. Second, there may be significant interactive effects between capital and OM&B projects that complicate separately estimating savings from these two sources.

Finally, another limitation of this approach is that it may be difficult or impossible to estimate the uncertainty of any OM&B savings estimate. Unless an estimate of uncertainty for the capital or retrofit project savings is available, evaluators will be unable to estimate the uncertainty of the OM&B savings, as the uncertainty of the OM&B savings depend on the uncertainty of both the regression-based SEM savings and capital project savings estimates. This protocol recommends against assuming capital project savings estimates have zero uncertainty.

3.8 Reporting Results

Evaluators should report point estimates of SEM program savings for the reporting period and standard errors or confidence intervals to indicate the program savings uncertainty. Depending on the evaluation objectives and research design, evaluators may also want to report savings estimates for individual facilities. Savings should be reported in units of energy and in a percentage of the adjusted baseline. Important aspects of the savings estimation should be clearly documented, as described in the preceding section addressing documentation.

4 Measurement and Verification Methods

This protocol recommends statistical analysis of facility energy consumption for estimating SEM program savings. This section provides guidance about specific estimation methods. It first describes and illustrates five different regression-based methods for estimating savings, followed by a discussion of non-routine adjustments to facility energy consumption and onsite data collection.

This section is technical in nature. It uses mathematical notation and applies basic statistical and econometric concepts to define key concepts and present the savings estimation methods. Since some readers may find the presentation challenging, numerical examples are included to demonstrate the application and facilitate understanding of key concepts and methods.

4.1 Regression and Savings Estimation Methods

This section presents five regression-based methods for estimating SEM savings:

- Forecast models
- Pre-post models
- Normal operating conditions models
- Backcast models
- Panel models.

All of the methods are based on Option C of the IPMVP, as each uses regression to adjust the baseline for differences in facility operating conditions between the baseline and reporting periods. The forecast method and the pre-post method are the most widely used by SEM program evaluators. All of the methods are expected to yield unbiased estimates of SEM savings if the energy consumption models accurately represent true facility energy consumption and the standard regression assumptions hold. This document's appendix proves the forecast and pre-post methods produce unbiased SEM-savings estimates under standard assumptions.

To make the presentation of the models concrete, suppose an industrial or large commercial facility participates in a ratepayer-funded SEM program. An evaluator wishes to estimate the facility savings during the program reporting period. The evaluator collects data on energy consumption for each of the T time intervals of the baseline period and each of the T^P time intervals of the reporting period. For example, the evaluator may collect facility energy consumption data for 24 months of the baseline period and 12 months of the reporting period.

The evaluator also collects interval data on the principal factors affecting energy consumption at the facility during the baseline and reporting periods.

Suppose that the evaluator determines that facility energy consumption in interval t, e_t, should be modeled as follows:

$$e_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_K x_{Kt} + \varepsilon_t \qquad \text{Equation 1}$$

where:

β_k = Coefficient to be estimated indicating effect of variable x_k on energy consumption.

x_{kt} = Variable k, k=1, 2, …, K, affecting facility energy consumption in interval t. For example, for an industrial facility, x_1 might be a measure of facility output, x_2 might be an indicator variable for facility closures, and x_3 might be a variable for outside temperature.

ε_t = Model error for energy consumption in interval t. The error term, ε_t, is assumed to be normally, independently, and identically distributed with mean zero and variance σ^2.

4.1.1 Forecast Models

With forecasting, the evaluator estimates a facility energy consumption regression with baseline period data and then uses the estimated regression to predict what facility energy consumption would have been during the reporting period had the facility not implemented SEM. The evaluator then estimates savings by comparing this adjusted baseline with metered energy consumption.

Specifically, the first step is to estimate Equation 1 using baseline period data. Then for each interval during the reporting period, the evaluator uses the estimated coefficients of Equation 1, b_0, b_1,…, b_k , to predict the adjusted baseline:

$$\widehat{e_t^P} = b_0 + b_1\, x_{1t}^P + b_2\, x_{2t}^P + \ldots + b_K\, x_{Kt}^P \qquad \text{Equation 2}$$

where x_{kt}^P is the k^{th} explanatory variable for time interval t of the reporting period. Again, predicted energy consumption is an estimate of what energy consumption would have been had SEM not been implemented and other facility conditions during the baseline period persisted during the reporting period.

Energy savings during interval t of the reporting period, $\hat{s}_t$, is estimated as follows:

$$\hat{s}_t = \widehat{e_t^P} - e_t^P$$

Energy savings during the reporting period, S, equals the sum of savings over the T^P intervals:[33]

$$\hat{S} = \sum_{t=1}^{T^P} \widehat{s_t}$$

The evaluator can estimate the variance and standard error of the forecast model savings estimate using standard regression software packages. As the appendix shows, the standard error of the forecast model savings estimate should be calculated as:

[33] By summing the estimated savings over appropriate time intervals, the evaluator can estimate savings for different periods, such as for the first or second year of an SEM program.

$$\text{standard error}(\hat{S}) = \sqrt{Var(b_0 * T^P + b_1 \sum x_{1t}^P + b_2 \sum x_{2t}^P + \ldots + b_k \sum x_{kt}^P) + T^P \widehat{\sigma^2}}$$

where:

$\widehat{\sigma^2}$ = The regression standard error; that is, the estimate of the error variance σ^2 from the baseline period regression model.

The first term in the formula is the variance of the adjusted baseline. It can be obtained using standard statistical software by expressing the sum of the interval adjusted baseline consumption as a linear combination of the estimated coefficients, where the factor multiplying each coefficient is the sum of the independent variable over the reporting period intervals. Specifically, evaluators should rewrite SEM savings as follows:

$$\hat{S} = \sum_{t=1}^{T^P} \widehat{s_t}$$

$$= \sum_{t=1}^{T^P} b_0 + b_1 x_{1t}^P + b_2 x_{2t}^P + \ldots + b_k x_{kt}^P - e_t^P$$

$$= b_0 * T^P + b_1 \Sigma x_{1t}^P + b_2 \Sigma x_{2t}^P + \ldots + b_k \Sigma x_{kt}^P - \Sigma e_t^P$$

where, again, each sum is taken over the intervals of the reporting period.

In a statistical software package (e.g., SAS, Stata, R), the evaluator needs to invoke a post-model estimation command to estimate the variance of this linear combination of coefficients.[34]

The second term in the standard error formula, $T^P \widehat{\sigma^2}$, is an estimate of the variance of the metered energy consumption during the reporting period. It may be estimated using the regression standard error (i.e., the regression root mean square error) of the baseline regression, under the assumption that the error variance during the baseline and reporting periods are equal.

4.1.1.1 Example of Forecast Model Savings Estimation

The following example illustrates the application of the forecast approach for estimating SEM program facility savings.

Table 1 shows monthly observations of average daily electricity consumption and output for a hypothetical industrial facility. The first 24 months correspond to the baseline period and the last 12 months correspond to the SEM reporting period.

[34] In SAS, the evaluator can use the *estimate* command in Proc GLM. In Stata, the evaluator can invoke the post-estimation command *lincom*. In R, the evaluator can use either the coef() or summary() functions on an lm() or glm() model object.

Table 1. Example Industrial Facility Energy Consumption and Output Data

Month	Average daily consumption (kWh)	Average daily output (units)	SEM	Month	Average daily consumption (kWh)	Average daily output (units)	SEM
1	8,164	23.9	0	19	6,318	15.7	0
2	7,352	20.1	0	20	6,505	14.5	0
3	6,869	19.2	0	21	7,481	20.2	0
4	6,429	16.0	0	22	7,653	23.7	0
5	5,815	13.2	0	23	6,422	15.4	0
6	6,578	18.1	0	24	7,271	21.3	0
7	7,889	23.3	0	25	5,201	12.0	1
8	5,439	11.6	0	26	5,669	21.8	1
9	6,049	11.5	0	27	4,312	19.9	1
10	6,266	13.5	0	28	2,951	11.6	1
11	5,898	12.0	0	29	3,520	19.7	1
12	6,801	17.6	0	30	4,704	24.8	1
13	6,654	19.4	0	31	2,416	8.6	1
14	6,097	14.0	0	32	3,669	15.3	1
15	7,215	21.5	0	33	3,270	15.3	1
16	7,387	20.1	0	34	3,909	21.1	1
17	5,641	13.2	0	35	4,584	24.7	1
18	7,394	20.8	0	36	3,710	18.4	1

Data source: Simulated by the authors using the following energy consumption model: average daily kWh = 4010 + 155*Average Daily Output – 2005 * SEM – 62*SEM*Average Daily Output + ε, where $\varepsilon \sim N(0, \sqrt{200})$. SEM savings ramped up in increments of 25% over the first four program months.

Figure 2 plots the output and energy consumption, showing that both appear to be highly correlated. Also, a reduction in energy consumption is evident after month 25, which coincides with the beginning of SEM implementation.

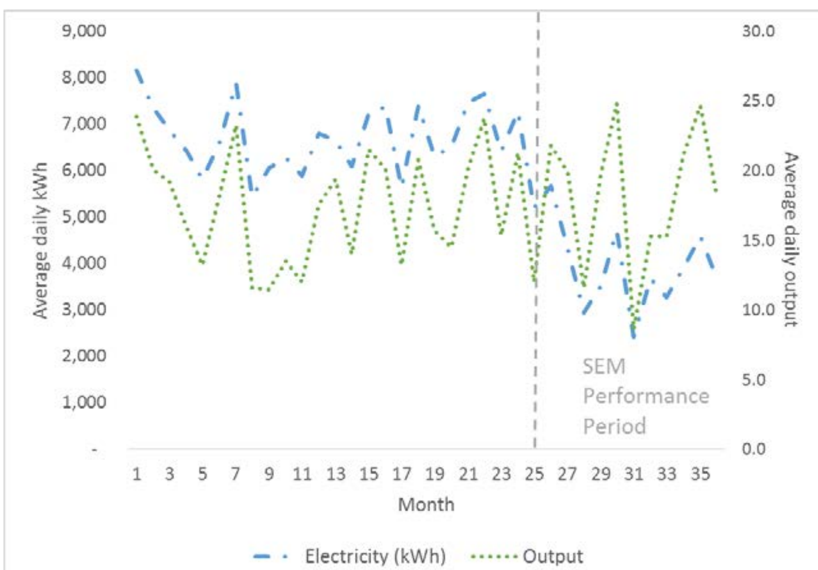


Figure 2. Plot of SEM facility electricity consumption and output vs. time

Suppose that, using the model-selection methods described in Section 3 of this protocol, the evaluator posits the following regression model of facility kWh during the baseline period:

$$kWh_t = \beta_0 + \beta_1 y_t + \varepsilon_t \qquad \text{Equation 3}$$

where:

kWh_t = Facility average daily electricity consumption in month *t*.

β_0 = Constant term to be estimated, indicating average daily electricity consumption unrelated to the facility output.

β_1 = Coefficient to be estimated, indicating the effect of an additional unit of output on electricity consumption.

y_t = Facility average daily output in month *t*.

ε_t = Error term.

The evaluator estimates the model using the first 24 monthly observations from the baseline period data.

Table 2 shows results from the OLS estimation of Equation 3. The model coefficients are estimated precisely—each is statistically significant at the 1% level—and have the expected signs. The coefficient on average daily output indicates average energy consumption increased by an average of 176 kWh for each unit of output.

Table 2. Estimates of Facility Forecast Regression Model

Dependent Variable	Average Daily kWh
Intercept	3,653* (214.4)
Average daily output	176.1* (12.0)
Regression Standard Error	229.06
F statistic	216.7
R^2	0.908
N	24

Note: Model estimated by OLS. Standard errors in parentheses.
* Denotes statistically significant at the 1% level.

Next, using the regression results, the evaluator estimates the adjusted baseline for each month of the reporting period. Monthly adjusted baseline electricity consumption (kWh) equals (3,653 + 176*average daily output during the month) times the number of days in the month.

Table 3 shows the calculation of the monthly adjusted baseline and SEM savings. Monthly SEM savings were estimated as the difference between the adjusted baseline and metered energy consumption.

Table 3. Estimates of Facility Adjusted Baseline Energy Consumption and Savings

SEM reporting period month	Average daily output (units)	Metered average daily consumption (kWh)	Adjusted baseline average daily consumption (kWh)	Days	SEM monthly savings	Cumulative SEM savings to date
25	12.0	5,201	5,772	31	17,698	17,698
26	21.8	5,669	7,499	28	51,235	68,933
27	19.9	4,312	7,152	31	88,059	156,992
28	11.6	2,951	5,693	30	82,264	239,257
29	19.7	3,520	7,121	31	111,640	350,897
30	24.8	4,704	8,023	30	99,573	450,470
31	8.6	2,416	5,175	31	85,524	535,994
32	15.3	3,669	6,343	31	82,884	618,878
33	15.3	3,270	6,338	30	92,041	710,919
34	21.1	3,909	7,377	31	107,480	818,399
35	24.7	4,584	7,994	30	102,291	920,690
36	18.4	3,710	6,891	31	98,608	1,019,298

Note: For description of calculations, see text. -

Lastly, the evaluator estimates savings for the first SEM program year by summing the monthly SEM savings for the first 12 reporting period months.

The last column of Table 3 shows the cumulative savings to date. By the end of the first year, it is estimated that the program had saved approximately 1,019,000 kWh. Based on implementation of Equation 2, the standard error of the savings estimate is 17,646 kWh and the estimated 95% confidence interval for the SEM savings is [984,710 kWh, 1,053,885 kWh].

4.1.2 Pre-Post Models

An alternative to the forecast approach is to use baseline period and reporting period data to estimate the facility average energy savings per time interval as a parameter of the regression model. This pre-post modeling approach estimates a modified version of Equation 1, with additional variable(s) to indicate the occurrence of SEM activity:

$$e_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + \theta d_t + \varepsilon_t \qquad \text{Equation 4}$$

where:

d_t = An indicator variable for SEM activity at the facility. It equals one if the facility initiated SEM in the current or in a previous interval; it equals zero otherwise.

θ = A coefficient to indicate the average effect per time interval of SEM activity on facility energy consumption.

The main difference between this model and the forecast model is that the pre-post model is estimated using both baseline period and reporting period data. The pre-post model also includes an indicator variable d_t to signify SEM program activity. A third difference is that the forecast model does not make any assumptions about how savings depend on the model explanatory variables. In contrast, savings are assumed to have a "level effect" on energy consumption for this pre-post model. Since d_t enters the model without being interacted with any other variables, savings do not depend on any of the independent variables in Equation 4.

Energy savings equal the product of the facility average savings per time interval and the number of time intervals during the reporting period:

$$S = \theta T^P$$

If $\hat{\theta}$ is the estimate of θ, then the variance of the estimated savings $\hat{S}$ equals:

$$\mathrm{var}(\hat{S}) = \mathrm{var}(\hat{\theta})\ (T^P)^2$$

4.1.2.1 Estimating SEM Savings in Multiple Sub-Periods

Evaluators may want to estimate savings for multiple periods to obtain savings estimates for different program years or to track growth, persistence, or decay of savings over time. To estimate SEM savings in multiple reporting periods, the evaluator can add more SEM reporting-period indicator variables to the regression, as follows:

$$e_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \dots + \beta_k x_{kt} + \sum_{j=1}^{J} \theta_j d_{jt} + \varepsilon_t \qquad \text{Equation 5}$$

where:

$d_{j,t}$ = An indicator for SEM activity in sub-period j, j = 1, 2, …, J, of the reporting period. This variable equals one if time interval t is in the j^{th} sub-period and the facility implemented SEM in the current interval or a previous interval; it equals zero otherwise.

θ_j = A coefficient indicating SEM average energy savings per interval during the j^{th} sub-period. The interval savings are measured relative to the baseline period.

As an objective of the SEM programs is continuous improvement of energy efficiency, evaluators may want to measure year-over-year changes in savings. Evaluators can use Equation 5 to measure these changes. Suppose that the time intervals are days and $d_{j,t}$ is an indicator variable for the jth program year. Then the incremental annual energy savings between the second and third program years would be calculated as follows:

$$\text{Incremental annual savings}_{Yr2,Yr3} = 365*(\theta_3 - \theta_2)$$

The incremental annual savings between other program years can be estimated analogously.

4.1.2.2 Estimating SEM Savings as a Function of Output or Weather

Equation 3 assumes that SEM resulted in a level-shift in facility energy consumption. In other words, the SEM's impact did not depend on output, weather, occupancy, or other variables affecting the facility's energy consumption. This might be a reasonable assumption for facilities where savings from SEM improvements did not vary closely with output or other variables. For example, a facility undertaking a lighting retrofit might have savings that do not vary with the facility's output. In many facilities, however, SEM savings will closely correlate with output or other observed drivers of energy consumption, such as occupancy. In this case, the evaluator can model SEM savings as a function of the model-independent variables:

$$e_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + \theta d_t + \theta_k d_t * x_{kt} + \varepsilon_t \qquad \text{Equation 6}$$

where all variables are defined as before, except:

θ_k = A coefficient indicating the SEM average energy savings, per time interval, per unit change of variable x_k.

In this specification, SEM can have a level savings effect, indicated by θ, as well as a slope-shift savings effect that depends on the variable x_k. For example, if variable x_k is facility output, then θ_k is the SEM savings per unit of output.

Energy savings during the reporting period would equal:

$$S = \theta T^P + \theta_k \sum_{t=1}^{T^P} x_{k,t}$$

4.1.2.3 Example of Pre-Post Regression Model Savings Estimation

This section illustrates a pre-post regression savings estimation, using data for all 36 intervals from the baseline and reporting periods in Table 1.

Again, in this example the evaluator wishes to estimate savings for the first SEM program year, thus specifying the following pre-post model:

$$kWh_t = \beta_0 + \beta_1 y_t + \theta d_t + \theta_1 y_t * d_t + \varepsilon_t \qquad \text{Equation 7}$$

where:

kWh_t = Facility average daily energy consumption in month t.

β_0 = Coefficient to be estimated, indicating facility average daily electricity consumption during the baseline period.

β_1 = Coefficient to be estimated, indicating average facility electricity consumption per unit of output.

y_t = Facility average daily production output during month t.

θ = Coefficient to be estimated, indicating SEM average electricity savings per day for the facility's baseload. These are savings from energy consumption that do not vary with the amount of output.

d_t = Indicator variable for SEM program activity. This variable equals one if SEM was implemented in the current month or in a previous month; it equals zero otherwise.

θ_1 = Coefficient to be estimated, indicating SEM average electricity savings per unit of output.

ε_t = Model error.

This specification includes an indicator variable for SEM activity, as well as for the SEM indicator interacted with output. The evaluator includes both variables with the expectation that SEM has both level and per-unit-of-output effects on facility energy consumption.

Table 4 shows estimates of the coefficients presented in Equation 7. The first column shows estimates of the coefficients in Equation 7. The second column shows estimates of Equation 7 without the interaction variable between the SEM indicator and output (to demonstrate the effect on estimated savings of misspecifying the energy consumption model).

Table 4. Pre-Post Regression Model Estimates

Dependent Variable	Pre-Post Model 1 Average daily kWh	Pre-Post Model 2 Average daily kWh
Intercept	3,652.9*** (454.3)	4,208.2*** (355.9)
Average daily output	176.1*** (25.3)	144.3*** (19.5)
SEM	-1,536.2** (688.1)	-2,779.8*** (178.0)
SEM*Average daily consumption	-70.5* (37.8)	
F statistic	105.8	146.0
R^2	0.908	0.898
N	36	36

Notes: Output based on analysis of data in Table 1. Model estimated by OLS. Standard errors in parentheses.
*, **, *** denotes statistically significance at the 1%, 5%, and 10% levels, respectively.

According to Pre-Post Model 1, SEM reduced energy consumption by an average of about 1,536 kWh per day, plus approximately 71 kWh per unit of output. The SEM program coefficients were statistically significant at the 5% and 1% levels, respectively. Since output averaged 17.8 units per day across the reporting period, the SEM program averaged savings of 2,790 kWh per day (=17.8*70.5 + 1,536.2).

Though the second model was misspecified because it omitted the interaction between the SEM indicator variable and output, the second model yielded an estimate of savings very similar to that of the correctly-specified Model 1. According to Pre-Post Model 2, daily savings from SEM averaged 2,780 kWh. Nevertheless, Model 1 has the advantage of allowing electricity savings to be decomposed into baseload savings and savings per unit of output and, therefore, may yield more useful information to the evaluator or program implementer.

The evaluator can then use the pre-post regression model to obtain an estimate of SEM program annual savings. Using the results of Model 1, the evaluator estimates annual savings as the sum of energy savings from baseload and production energy consumption:

Annual SEM savings = days*1536.2 kWh/day + annual output*70.5 kWh/unit of output

Assuming the facility operated 365 days and that annual output equaled 6,467 units, annual SEM energy savings equaled 1,016,576 kWh. The estimated 95% confidence interval equaled [893,745 kWh, 1,139,407 kWh].[35]

These estimates can be compared to an annual savings estimate from the forecast Model 1 of 1,019,298 kWh. The pre-post Model 1 and Model 2 yielded estimates of annual savings of 1,016,576 kWh and 1,014,608 kWh, respectively.

4.1.3 Comparison of Forecast and Pre-Post Approaches

The forecast and pre-post models take different approaches to estimating savings. The forecast approach fits a model using data from the baseline period and then uses that model to predict energy consumption in the reporting period. The pre-post approach fits one model with SEM level-shift or slope-shift indicator variables using data for the baseline and reporting periods.

Despite these differences, the forecast and pre-post models are expected to yield similar estimates of the adjusted baseline and SEM savings, as illustrated in the preceding comparison of the forecast and pre-post model savings estimation examples. The equivalence of the two approaches is analyzed from a conceptual perspective in this protocol's appendix. The models yielded the same predictions of the adjusted baseline, shown by identical intercepts and coefficients on average daily output for the two models. The models also yielded very similar savings estimates.

In general, as demonstrate in the appendix, the forecast and pre-post models produce unbiased savings estimates if the following two conditions hold:

(1) The pre-post model is specified as if SEM affects all energy consumption relationships modeled during the baseline period. Any variable expected to affect baseline period energy consumption should be interacted with an indicator variable for SEM and included in the regression.

In the above example, the pre-post model includes both an intercept for the reporting period (the SEM level shift) and an interaction between output and SEM (the SEM slope shift), thereby allowing baseload energy consumption and energy consumption per unit of output to differ between the baseline and reporting periods:

[35] The confidence interval requires accounting for the covariance between the estimated coefficients on SEM and SEM*average daily consumption. The evaluator can calculate the confidence interval by outputting the variance-covariance matrix or by using statistical software such as SAS, STATA, or R.

(2) The forecast and pre-post models are correctly specified in the sense that the energy consumption regression models closely approximate the facility's true energy consumption relationships during the baseline period. The models do not omit variables that were correlated with SEM implementation and facility energy consumption.

In this protocol's examples, the true energy consumption relationships are known because the data are simulated. In general, however, the evaluator will not know the true facility energy consumption model and the forecast and pre-post models may produce biased savings estimates. To obtain a valid savings estimate, the evaluator should collect facility data to build a valid model of facility energy consumption. Section 3 of this protocol describes the data collection and model specification processes for SEM evaluation.

4.1.4 Normalized Operating Conditions Models

The forecast and pre-post models produce estimates of SEM energy savings for the reporting period. The savings reflect the facility's operating conditions during the reporting period. However, operating conditions during the reporting period may have been atypical, producing savings that the facility may not expect in most years. Instead, evaluators may want an estimate of annual savings for the facility under normal operating conditions, which might be characterized by particular expected weather, occupancy levels, or production.

Suppose that facility energy consumption for interval *t* of the baseline period, e_t, can be modeled as:

$$e_t = \beta_0 + \beta_1 x_t + \varepsilon_t \qquad \text{Equation 8}$$

and suppose that the facility's energy consumption for interval t of the reporting period, e_t^P, can be modeled as:

$$e_t^P = \beta_0^P + \beta_1^P x_t^P + \varepsilon_t^P \qquad \text{Equation 9}$$

where *P* denotes the reporting period and x_t is units of facility output, a weather-related variable, or occupancy. The beta coefficients, β_0 and β_1, indicate, respectively, the facility's baseload consumption per interval and the marginal effect of x_t on energy consumption. The beta coefficients for the reporting period, β_0^P and β_1^P, reflect any SEM impacts.

Furthermore, suppose that x_k^N is the normal or expected value of x for interval k, k=1, 2, …, K, of the calendar year. For example, x could be heating degrees and $x_1^N, x_2^N, \dots, x_K^N$ would be expected values of heating degrees for intervals (e.g., days, weeks, or months) of the calendar year.

Evaluators can obtain an estimate of SEM savings under normal operating conditions by following these steps:

(1) Estimate Equation 8, the facility consumption model for the baseline period, using baseline period data, and Equation 9, the facility consumption model for the reporting period, using reporting period data.

(2) Predict energy consumption under normal operating conditions for the baseline period and reporting period using estimates from Step 1 to obtain the normalized adjusted consumption for each interval k of the calendar year:

$$\widehat{e_k^N} = b_0 + b_1\, x_k^N \qquad \text{Equation 10}$$

$$\widehat{e_k^{N,P}} = b_0^P + b_1^P\, x_k^N \qquad \text{Equation 11}$$

(3) Estimate annualized energy savings under normal operating conditions, S^N, as the difference between normalized adjusted consumption for the baseline period and the normalized adjusted consumption for the reporting period.

$$S^N = \sum_{k=1}^{K} \widehat{e_k^N} - \sum_{k=1}^{K} \widehat{e_k^{N,P}} \qquad \text{Equation 12}$$

4.1.4.1 Example of Normalized Operating Conditions Savings Estimation

In Table 1, the industrial facility produced 6,497 units of output during the 12 months of the reporting period. Suppose that this output was abnormally low and that the facility usually produces 10,000 units of output annually. How much electricity would the facility save under normal operating conditions?

First, the evaluator would estimate the facility's electricity consumption during a normal year before implementing SEM. This can be calculated with the forecast model coefficients in Table 2. The facility would have consumed 3,094,345 kWh during a normal year before implementing SEM. This estimate was obtained as follows:

3,653.0 kWh/day*365 days + 10,000 units of output annually*176.1 kWh/unit of output

Next, using observations for months 25 to 36 of Table 1, the evaluator would estimate a consumption model for the reporting period. Table 5 shows the coefficient estimates from that regression.[36]

Table 5. Reporting Period Regression Model Estimates

Dependent Variable	Average Daily kWh
Intercept	2116.7** (850.7)
Average daily output	105.6** (46.1)
Regression Standard Error	799.0
F statistic	5.3
R^2	0.344
N	12

Notes: Model estimated by OLS. Standard errors in parentheses.
** Denotes statistically significant at the 5% level.

36 This example is illustrative only. The reader should keep in mind that 12 data points is a small number for estimating the reporting period regression and would want to exercise caution in a similar situation.

According to the model coefficients in Table 5, the facility would have consumed 1,828,682 kWh during a normal year after implementing SEM. This estimate was obtained as follows:

$$2{,}116.7 \text{ kWh/day} * 365 \text{ days} + 10{,}000 \text{ units annually} * 105.6 \text{ kWh/unit}$$

Taking the difference between the normalized adjusted consumption for the baseline and reporting period, the evaluator estimates that the facility can expect to save 1,265,663 kWh/year. Table 6 shows the normal operating conditions savings estimate.

Table 6. Normalized Operating Conditions Savings Estimate

	Annual kWh
Normalized Adjusted Consumption for Baseline Period (a)	3,094,345
Normalized Adjusted Consumption for Reporting Period (b)	1,828,682
Normalized Savings (a-b)	1,265,663

4.1.5 Backcast Models

Backcast modeling involves using reporting period consumption data to "backcast" consumption during the baseline period under reporting period conditions and then estimating SEM savings as the difference between the backcasted adjusted baseline and metered consumption. The backcast adjusted baseline represents facility consumption that would have occurred during the baseline period if the reporting period operating equipment and practices had been in place. As with any forecast method, this method requires developing a model that characterizes energy consumption as a function of relevant variables.

Evaluators may find the backcast approach useful when:

- There is limited data on energy consumption and corresponding independent variables during the baseline period but detailed data for the reporting period.
- Facility operating conditions during the reporting period are inclusive of facility operating conditions during the baseline period conditions, but not vice-versa.

For example, an industrial facility may have produced only low levels of output during the baseline period but low and high levels during the reporting period. A forecast model may produce an inaccurate estimate of adjusted baseline consumption because some reporting period conditions (i.e., high output levels) were outside of those experienced during the baseline. In contrast, the backcast adjustment approach is expected to yield valid predictions of baseline period energy consumption because the reporting period included low levels of output.

Evaluators should apply the backcast approach judiciously, considering whether the approach yields the desired savings estimate. Typically, evaluators will want an estimate of savings for the reporting period or for standard operating conditions. However, the backcast approach yields an estimate of counterfactual savings, what SEM energy savings would have been during the baseline period. If the facility's operating conditions differ substantially between the baseline and reporting periods, the backcast approach may not produce the desired estimate.

4.1.5.1 Example of Backcast Savings Estimation

Suppose an evaluator wanted to apply the backcast approach to the facility consumption data in Table 1. The evaluator first estimates a regression model of facility consumption using reporting period data for months 25–36. Table 5 shows results of that regression.

Next, the evaluator would use the regression coefficients in Table 5 to backcast the facility's consumption during the baseline period. Table 6 shows the facility would have consumed 1,414,907 kWh and 1,479,539 kWh during months 1–12 and months 13–24, respectively, of the baseline period if it had implemented an SEM.

The evaluator would then compare the backcasted adjusted baseline consumption with the metered consumption to estimate the backcast savings for the two baseline periods.

Table 7 presents the backcast estimates.

Table 7. Backcast Model Savings Estimates

	Months (1-12)	Months (13-24)
Baseline Period Consumption (kWh)	2,419,031	2,496,205
Backcast Adjusted Baseline Consumption (kWh)	1,414,907	1,479,539
Backcast Electricity Savings Estimate (kWh)	1,004,125	1,016,666

The evaluator should keep in mind that the backcast savings are estimates of counterfactual SEM savings during the baseline period. The backcast savings may not equal the actual savings the program achieved during the reporting period if other factors are substantially different. In this example, the backcast model produced annual savings estimates that were very close to the forecast model estimate of annual savings (1,019,000 kWh) because annual output levels during the baseline and reporting period were approximately equal. If output levels had differed, the forecast model and backcast model savings estimates would have differed, too.

4.1.6 Panel Regression Models

This protocol emphasizes analysis of individual facilities because many program administrators require an SEM-savings estimate for each facility. Also, many industrial and large commercial facilities have unique characteristics that make group analysis problematic. For example, food processors, lumber mills, hospitals, and wastewater treatment facilities have very different outputs, production processes, and energy-consumption characteristics. These differences make regression modeling for groups of very different facilities difficult.

There are, however, circumstances when group or panel analysis of energy consumption for a group of facilities may be appropriate. A panel consists of data for two or more facilities and multiple observations for each sampled facility. A panel dataset should cover the baseline and reporting periods. Panel regression analysis yields an estimate of the average savings per facility, per unit of time; this can provide a more economical means of program impact evaluation than estimating savings for each site.

Panel analysis is appropriate when the evaluator does not require facility-specific savings and when program populations or subpopulations have similar energy consumption characteristics.

For example, group analysis could be used to estimate SEM program average savings per facility for a population of office buildings or primary or secondary schools. These types of buildings have relatively similar energy end uses (lighting and space conditioning) and energy consumption intensities. This analysis also may be appropriate for an SEM program targeting a specific industrial sector, such as food processing.

Suppose the evaluator has data on baseline and reporting period energy consumption and energy consumption drivers for i=1, 2, …, N facilities.[37] Then a panel regression of facility *i* energy consumption during time interval *t* would be:

$$e_{it} = \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \theta d_{it} + \alpha_i + \varepsilon_{it} \qquad \text{Equation 13}$$

where all of the variables affecting energy consumption of a facility have been indexed by *k,* k=1,2, …, k, and the other variables are defined as before; *i* indexes the facility, and *t* indexes the time period.

For example, x_{1it} is the variable x_1 (e.g., outside temperature) for facility *i* during time interval *t*, and α_i is the error term specific to facility *i* that does not vary over time.[38] Instead of using energy consumption as the dependent variable, evaluators may want to normalize the dependent variable by dividing it by the number of square feet or the number of units of output to account for differences between facilities in floor area or other variables affecting energy consumption.

The term α_i may or may not be correlated with the *x* variables and d_{it}. An evaluator who believes α_i is correlated should estimate a *fixed effects* model, which involves estimating Equation 8 by OLS, with a separate intercept for each facility in the analysis sample. The facility intercepts control for all unobservable, time-invariant factors specific to the facility that may be correlated with the other variables in the model.

Alternatively, an evaluator who believes α_i is uncorrelated with the independent variables should estimate a *random effects model,* which involves estimating Equation 8 by generalized least squares, first by estimating the covariance matrix of the error term, and then using the estimated covariance matrix in a second-stage estimation of the Equation 8.

In general, when there is a choice between the two estimation methods, fixed effects estimation is recommended because it yields consistent estimates of the model parameters when the

[37] This panel regression approach assumes that reference energy use was estimated using pre-SEM engagement facility energy use of SEM participants. An alternative approach for estimating reference energy use would be to identify a comparison group of nonparticipant facilities and to use their energy use during the SEM performance period as a baseline. See Agnew (2013) for baseline approaches employing a control group.

[38] The regression specification excludes time interval fixed effects, which would capture impacts of each time interval on average facility energy use. If there is no variation between facilities in the data of first SEM implementation, the evaluator will be unable to include both time interval fixed effects and an SEM indicator variable because the SEM indicator variable and the fixed effects will be co-linear. If the regression includes interaction variables between the SEM indicator and other variables but not an SEM indicator, the evaluator could include time interval fixed effects in the regression. If the number of facilities is sufficiently large and there is enough variation between facilities in the date of first SEM implementation, the evaluator can include time interval fixed effects.

assumptions of the random effects model or the fixed effects model hold true (Greene 2012). The random effects estimator, however, is not consistent when the assumptions of the fixed effects model hold true.

Estimation of Equation 8 yields an estimate of the average program effect θ. With the panel regression model, program savings can be estimated as shown:

$$S = \theta \sum_{i=1}^{N} \sum_{t=1}^{T^P} d_{i,t}$$

The program savings are the product of the average savings per facility, per interval, and the total number of facility SEM engagement intervals during the reporting period.

While panel regression analysis does not yield a savings estimate for each facility, it can be used to estimate how program effects depend on the preexisting characteristics of participants. For example, the model can be used to estimate savings as a function of floor area or by school type (e.g., elementary, secondary). Evaluators can do this by interacting indicators for program activity with participant characteristic variables.

4.2 Non-Routine Adjustments

Evaluators may need to make *non-routine adjustments* to improve the accuracy of the adjusted baseline. A non-routine adjustment refers to a one-time, *ad hoc* adjustment to the adjusted baseline to account for a change in facility energy consumption that cannot be modeled econometrically. Not accounting for such changes may bias the savings estimate. Evaluators, however, should make these adjustments sparingly and objectively, without regard to the expected effect on the savings estimate.

For example, suppose an industrial facility replaced equipment and implemented SEM at the same time. The equipment replacement was scheduled far in advance of SEM implementation; however, both had the effect of reducing energy consumption per unit of output. Since the equipment replacement and SEM implementation coincided, the evaluator may not be able to use regression analysis to identify the SEM savings.

In such instances, if an engineering-based estimate of the change in energy consumption is available, the evaluator can adjust the adjusted baseline consumption to account for the equipment change. The difference between the regression and non-routine adjusted baseline and metered energy consumption would then yield an estimate of the SEM savings. If an estimate of the impact of the change in energy consumption is *not* available, it may not be feasible to use statistical methods to estimate the SEM savings.

Non-routine adjustments of this type should be used sparingly. The evaluator should first attempt to account for the change in energy consumption in the regression model. In the above example, if the equipment replacement had been a more efficient space conditioning system and SEM energy savings did not depend on weather, the evaluator might be able to use regression to control for the equipment replacement by modeling energy consumption as a function of HDDs, CDDs, and the date of the equipment change.

When non-routine adjustments must be made, evaluators should apply them based on careful engineering analysis, precisely documenting all assumptions and calculations. The evaluator should carefully review the assumptions and accuracy of the calculations.

4.3 Site Data Collection

Thus far, this protocol has assumed evaluators would not perform primary data collection; rather, that they would analyze data on facility energy consumption, output, and weather collected from program implementers, the utility, or third-party data providers. The exception would be conducting interviews with facility staff or program implementers to gather additional information about the facility's energy consumption and implementation of SEM. Such primary data collection can greatly improve evaluators' understanding of facility energy consumption, and this protocol highly recommends conducting these interviews.

In some circumstances; however, evaluators may be able to significantly improve the accuracy of SEM-savings estimates by conducting onsite facility inspections and data collection. Many SEM program facilities install capital equipment or retrofit measures as a result of SEM engagement. Other facilities may have installed capital measures during the baseline period.

Evaluators can use site visits to improve the accuracy of capital project savings estimates needed for developing a baseline model or estimating SEM savings. Specifically, site visits can verify key assumptions in the calculation of capital project savings. Evaluators also can use site visits to check the reasonableness of SEM-savings estimates obtained from statistical models.

More specifically, this protocol recommends evaluators consider conducting site visits when one or more of the following conditions hold true:

- An evaluation objective is to obtain separate estimates of SEM capital measure savings and SEM operations, maintenance, and behavioral savings;
- Savings from capital measures constitute a large share of SEM savings, and the statistical analysis yields an SEM-savings estimate with substantial uncertainty; or
- It is necessary to perform a one-time, non-routine adjustment to the baseline or reporting-period energy consumption to account for capital measure savings or for a change in facility operations, and a site visit can significantly reduce uncertainty about the magnitude of such adjustments.

When one or more of these conditions hold, an onsite M&V that better characterizes the impacts of such changes on facility energy consumption may improve the accuracy of the SEM-savings estimates.

This protocol recommends that evaluators follow IPMVP (2012), which recommends best practices for conducting onsite data collection for the evaluation of capital measure and retrofit projects. For capital equipment and retrofit measures installed as part of SEM engagement the most appropriate evaluation options are as follows:

- ***Operational Verification.*** For this type of savings estimation method, the evaluator relies on a variety of onsite data collection activities (e.g., visual inspections, spot measurements, data trending reviews) to verify an energy efficiency measure is installed and functioning as intended.

- ***IPMVP Option A, Retrofit Isolation: Key Parameter Measurement.*** For this method, the evaluator uses engineering calculations and partial site measurements to verify the savings resulting from specific measures. The evaluator estimates some parameters that are not measured.
- ***IPMVP Option B, Retrofit Isolation: All Parameter Measurement.*** For this method, the evaluator uses engineering calculations and ongoing site measurements to verify savings as the change in energy consumption of the affected system. This may be appropriate for variable frequency drives, where the evaluator could use long-term metering to determine the true reduction in motor energy over various seasonal and loading cycles.

Evaluators should know that IPMVP Option A and Option B typically require baseline- and reporting-period data, and that baseline-period data may be unavailable if not previously collected.

When selecting an onsite data gathering approach, the evaluator should seek to balance the expected reduction in uncertainty with the project's resources and budget. To decrease the uncertainty of estimates, the evaluator should measure and meter where experience has shown that energy consumption can vary widely. The evaluator also should measure and meter in situations where existing estimates of capital project savings remain uncertain. Through this approach, the evaluator can confirm that the reported capital and retrofit measures are (1) installed, (2) functioning, and (3) operating appropriately. If the evaluator determines that the results from an installed measure differ from the assumptions expected in the approach, additional data may be collected to further evaluate the energy savings.

5 Other Evaluation Issues

5.1 Sampling

Some SEM programs may enroll a large number of facilities; however, they have evaluation budgets too small to support an impact evaluation of the population of facilities. In this case, the evaluator may need to analyze a random sample of facilities from the program population.

Evaluators can consult well-known guidelines and protocols for simple random sampling, stratified random sampling, and other, more complex sampling designs for efficiency program populations. Evaluators can find useful sampling guidelines in *UMP Chapter 11: Sample Design Cross-Cutting Protocols* (Khawaja 2013). *Sampling Techniques* (Cochran 1977) provides another good reference.

5.2 Free-Ridership, Spillover, and Net Savings

This protocol is primarily concerned with estimation of SEM program gross savings using a regression-adjusted baseline. The issues and approach for estimating SEM net savings are very similar to those for other ratepayer-funded, energy efficiency measures. This protocol recommends that evaluators consult *UMP Chapter 23: Estimating Net Savings: Common Practices* (Violette 2014) for guidance.

6 References

Agnew, K.; Goldberg, M. (April 2013). Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. NREL.

Amundson, T.; et al. (2013). "Elements of Defensible Regression-Based Energy Models for Monitoring and Reporting Energy Savings in Industrial Energy Efficiency Operation and Maintenance Projects." Paper presented at the 2013 ACEEE Summer Study on Energy Efficiency Industry.

ASHRAE. (2014). *ASHRAE Guideline 14-2014—Measurement of Energy and Demand Savings. Atlanta, Ga.: American Society of Heating, Refrigerating and Air-Conditioning Engineers*.

Bonneville Power Administration. (2012). *Regression for M&V: Reference Guide*. Prepared for BPA by Research into Action, Inc. et al. https://www.bpa.gov/EE/Policy/Imanual/Documents/July%20documents/3_BPA_MV_Regression_Reference_Guide_May2012_FINAL.pdf

Cadmus Group. (February 2013). *Energy Smart Industrial—Energy Management Pilot Impact Evaluation.* Prepared for Bonneville Power Administration. http://www.bpa.gov/EE/Utility/research-archive/Documents/BPA_Energy_Management_Impact_Evaluation_Final_Report_with_Cover.pdf

Cochran, W. G. (1977). *Sampling Techniques.* New York: John Wiley & Sons.

Consortium for Energy Efficiency. (2014a). *CEE Strategic Energy Management Minimum Elements*. Available at: https://library.cee1.org/content/cee-strategic-energy-management-minimum-elements/

Consortium for Energy Efficiency. (2014b). *CEE Industrial Strategic Energy Management Initiative*. http://library.cee1.org/sites/default/files/library/11282/Industrial_SEM_Initiative.pdf

Consortium for Energy Efficiency. (2015). *SEM Program Case Studies Report.* http://library.cee1.org/content/sem-program-case-studies-2014

Consortium for Energy Efficiency. (2016). *Strategic Energy Management Initiative – 2015 CEE Annual Report.* https://2015annualreport.cee1.org/initiatives/strategic-energy-management-initiative/

Deru, M.; Torcellini, P. (2007). Source Energy and Emission Factors for Energy Use in Buildings. National Renewable Energy Laboratory Technical Report NREL/TP-550-38617. http://www.nrel.gov/docs/fy07osti/38617.pdf.

DNV KEMA Energy and Sustainability; Research Into Action, Inc. (April 2014). *NEEA Industrial Initiatives—Market Progress Evaluation Report #8*. Prepared for the Northwest Energy Efficiency Alliance. http://neea.org/docs/default-source/reports/neea-industrial-initiatives–market-progress-evaluation-report-8.pdf?sfvrsn=10

Efficiency Valuation Organization. (2012). *International Performance Measurement and Verification Protocol – Concepts and Options for Determining Energy and Water Savings Volume 1*. Washington, D.C.: Efficiency Valuation Organization.

Energy 350. (April 2014). *Small to Medium Industrial SEM Energy Savings Validation.* Prepared for the Northwest Energy Efficiency Alliance. http://neea.org/docs/default-source/reports/small-to-medium-industrial-sem-energy-savings-validation.pdf?sfvrsn=12

Energy Information Administration. (2008). Commercial Building Energy Consumption Survey, Table E1A, Major Fuel Consumption (Btu) by End Use for All Buildings (2003). http://www.eia.gov/consumption/commercial/data/2003/pdf/e1a-e11a.pdf

Frison, L.; Pocock, S.J. (1992). "Repeated Measures in Clinical Trial: Analysis Using Mean Summary Statistics and Its Implications from Design." Statistics in Medicine 11: 1685-1704.

Greene, W. H. (2012). *Econometric Analysis.* 7th International Edition. Essex, England: Pearson Education.

Imbens, G.; Rubin, D.M. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. New York: Cambridge University Press.

List, J.A.; Sadoff, S.; Wagner, M. (2010). "So You Want to Run an Experiment, now What? Some Simple Rules of Thumb for Optimal Experimental Design." National Bureau of Economic Research Paper 15701.

Khawaja, M. S.; Rushton, J.; Josh Keeling J. (April 2013). *Chapter 11: Sample Design Cross-Cutting Protocols.* The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. NREL.

Navigant Consulting, Inc. (2013). *Impact Evaluation of Energy Trust of Oregon's 2009–2011 Production Efficiency Program.* December 2013. Prepared for the Energy Trust of Oregon. http://assets.energytrust.org/api/assets/reports/PE_Impact_Eval_2009-11.pdf

Northwest Industrial Strategic Energy Management Collaborative. (2013a). *Tools and Methods for Addressing Auto-correlation in Energy Modeling.* https://conduitnw.org/Pages/File.aspx?rid=1763

Northwest Industrial Strategic Energy Management Collaborative. (2013b). *Tools and Methods for Addressing Multicollinearity in Energy Modeling.* https://conduitnw.org/pages/file.aspx?rid=1762

Price, P.N. (2014). Review of Prior Commercial Building Energy Efficiency Retrofit Evaluation: A Report to Snohomish Public Utilities District. Ernest Orland Lawrence Berkeley National Laboratory. https://publications.lbl.gov/islandora/object/ir%3A1005583/datastream/PDF/view

Reddy, T.A.; Claridge, D.E. (2000). Uncertainty of 'Measured' Energy Savings from Statistical Baseline Models. HVAC & R Research: 6 (1), pp. 3-20.

Stern, F. (April 2013). *Chapter 10. Peak Demand and Time Differentiated Energy Savings Cross Cutting Protocols*. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. NREL.

Thiel, H. (1971). Principles of Econometrics. New York: John Wiley & Sons.

Violette D. M.; Rathburn, P. (September 2014). *Chapter 23: Estimating Net Savings: Common Practices*. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. NREL.

U.S. Department of Energy (2017a). *Qualified Energy Savings Measurement and Verification Protocol for Industry*.

U.S. Department of Energy (2017b). *Superior Energy Performance Measurement and Verification Protocol for Industry*.

U.S. Department of Energy (2016b). *Superior Energy Performance Certification Protocol for Industry*.

Appendix A

This appendix demonstrates the equivalence of the forecast savings and pre-post model approaches, showing that both produce unbiased savings estimates.

The appendix also derives the analytic formula for estimating the forecast savings standard error. The analytic formula captures two sources of uncertainty: (1) the variance of the estimated baseline model coefficients and (2) the variance of metered energy consumption during the reporting period. It is necessary to account for both components to obtain an accurate estimate of the forecast savings standard error.

The first appendix section presents a model of facility energy consumption and defines SEM savings. The second section proves that, under the assumptions of the classical linear regression model, the pre-post and the forecast savings estimation methods yield unbiased estimates of SEM savings. The third section derives the formula for the forecast model standard error.

A.1 Definition of SEM Savings

This section presents a general, or theoretical, overview of calculating SEM savings. The formulas developed in A.1 should not be used to actually calculate energy savings. Instead they are provided as reference to aid in demonstrating the equivalence of forecasting and pre-post modeling techniques in Sections A.2 and A.3.

Suppose the following regression model describes facility electricity consumption in the baseline period:

$$kWh_t = \alpha + \beta x_t + \varepsilon_t \quad \text{Equation 14}$$

where x_t is an explanatory variable (e.g., output) and α and β are coefficients to be estimated. α can be interpreted as baseload energy consumption per interval, and β can be interpreted as the energy consumption per unit of output. The error term ε_t is normally, independently, and identically distributed with mean zero and variance σ^2.

During the SEM reporting period, the facility implements changes to improve the efficiency of baseload energy consumption and energy consumption per unit of output. kWh_t^P is metered energy consumption during the baseline period; kWh_t^P can be expressed as the sum of the expected value of kWh_t^P, conditional on x_t^P plus an error:

$$kWh_t^P = E[kWh_t \mid x_t^P, \alpha^P, \beta^P] + \varepsilon_t^P$$

After implementation, facility electricity consumption during the SEM reporting period (P) is calculated as follows:

$$kWh_t^P = \alpha^P + \beta^P x_t^P + \varepsilon_t^P \quad \text{Equation 15}$$

where P denotes reporting period, kWh_t^P and x_t^P are energy consumption and output, and α^P and β^P are coefficients to be estimated. Baseload energy consumption per interval is α^P, and β^P is energy consumption per unit of output after implementation of SEM. The error term ε_t^P is normally, independently, and identically distributed with mean zero and variance σ^2_P. The variance of ε_t and ε_t^P may differ.

For interval *t* of the reporting period with facility output x_t, SEM energy savings s_t equals the difference between expected energy consumption, conditional on x_t^P under baseline conditions, and expected energy consumption, conditional on x_t^P under reporting period conditions:

$$s_t = E[kWh_t | x_t^P, \alpha, \beta] - E[kWh_t | x_t^P, \alpha^P, \beta^P]$$

$$= \alpha + \beta x_t^P - \alpha^P - \beta^P x_t^P$$

$$= (\alpha - \alpha^P) + (\beta - \beta^P) * x_t$$

where E is the expectation operator and | denotes "conditional on."

The first term is the baseline energy savings per interval, and the second term is the energy savings per unit of output, multiplied by the amount of output in interval *t*.

Savings for the reporting period with T intervals, denoted, t=1, 2, …, T^P equals:

$$S = (\alpha - \alpha^P)*T^P + (\beta - \beta^P) * \sum_{t=1}^{T^P} x_t$$

$$= \alpha^\Delta * T^P + \beta^\Delta * \sum_{t=1}^{T^P} x_t$$

where:

$$\alpha^\Delta = \alpha - \alpha^P\text{; and}$$

$$\beta^\Delta = \beta - \beta^P$$

A.2 Equivalency of Pre-Post and Forecast Savings Methods

The reporting period energy savings S can be estimated using the pre-post method or the forecast method. This section shows that the pre-post and forecast methods both yield unbiased estimates of S.

A.2.1 Pre-Post Method

The first approach nests both Equation 14 and Equation 15 in a single model, thereby obtaining the pre-post model; and then estimates the coefficients of the pre-post model:

$$kWh_t = \text{Baseline Energy Consumption} - \text{Savings} + \text{Error}$$

where:

$$\text{Baseline Energy Consumption} = \alpha + \beta x_t$$

$$\text{Savings} = \alpha^\Delta * Post_t + \beta^\Delta x_t * Post_t$$

Error $= \varepsilon_t + (\varepsilon_t^P - \varepsilon_t)*Post_t$

$Post_t$ = 1 for intervals during the reporting period and = 0, otherwise.

$$kWh_t = \alpha - \alpha^\Delta * Post_t + \beta x_t - \beta^\Delta x_t * Post_t + \varepsilon_t + (\varepsilon_t^P - \varepsilon_t)*Post_t \quad \text{Equation 16}$$

Note that if Post=0, the model reduces to Equation A.1, and, if Post=1, the model reduces to Equation A.2.

The model is estimated by OLS, producing an estimate of savings for interval *t*:

$$\widehat{s_t} = a^\Delta + b^\Delta xt$$

Reporting period savings equals the following $\widehat{S_t}$:

$$\widehat{S} = T^p * a^\Delta + b^\Delta * \sum_{t=1}^{T^P} x_t$$

Where a^Δ and b^Δ are the OLS, unbiased estimates of α^Δ and β^Δ, respectively.

Under the assumptions of Equation 14 and Equation 15, OLS will yield unbiased estimates of α, α^Δ, β, and β^Δ; therefore $\widehat{S}$ is an unbiased estimate of S.

A.2.2 Forecast Method

A second approach for estimating savings is the forecast method. Using data from t=1, 2, …, T periods during the baseline period, the researcher estimates Equation 14 by OLS and obtains estimates of α, β, and error variance σ^2, denoted a, b, and $\widehat{\sigma^2}$.[39]

Next, the researcher uses the model $\widehat{kWh_t} = a + b\,x_t$ to predict *expected energy consumption* in the reporting period (P), under the assumption that SEM had not been implemented. For each of the t=1, 2, … , T^P intervals during the reporting period, the researcher observes both kWh_t^P and x_t^P.

Energy savings in interval t of the reporting period are estimated as follows:

$$\widehat{s_t} = \widehat{kWh_t^P} - kWh_t^P$$

$$= a + bx_t^P - kWh_t^P$$

$$= a + b\,x_t^P - \alpha^P - \beta^P x_t^P - \varepsilon_t^P$$

where $\widehat{kWh_t^P}$ is an estimate of the expected energy consumption under baseline conditions during the reporting period (the forecast adjusted baseline), and kWh_t^P is metered energy consumption during the baseline period. In accordance with Equation 15, kWh_t^P can be expressed as the sum of the expected value of kWh_t^P, conditional on x_t^P plus an error; that is:

[39] Let e_t be the residual of the regression in period t. $\widehat{\sigma^2}$ is estimated as the sum of squared residuals, divided by T-k; that is, $\Sigma_{t=1}^{T} e_t^2/(T-k)$, where k is the number of coefficients to be estimated in the regression.

$kWh_t^P = E[kWh_t | x_t^P, \alpha^P, \beta^P] + \varepsilon_t^P$

This protocol uses this fact in calculating the variance of forecast savings (below).

Reporting period savings equals the following:

$$\hat{S} = \sum_{t=1}^{T^P} \hat{s_t}$$

$$= \sum_{t=1}^{T^P} a + b\, x_t^P - \alpha^P - \beta^P x_t^P - \varepsilon_t^P \qquad \text{Equation 17}$$

Taking expectations (E[ ]) of both sides of Equation 17:

$$E[\hat{S}] = (\alpha - \alpha^P) * T^P + (\beta - \beta^P) * \sum_{t=1}^{T^P} x_t^P$$

$$= \alpha^\Delta * T^P + \beta^\Delta * \sum_{t=1}^{T^P} x_t^P$$

The first equality follows because, under the assumptions of Equation 14, OLS yields an unbiased estimate of the model parameters: $E[a] = \alpha$ and $E[b] = \beta$. Therefore, $\hat{S}$ is an unbiased estimate of pilot savings, and the forecast method and the pre-post method are expected to provide unbiased estimates of S.[40]

A.3 Standard Error of Forecast Method Savings

This section first derives the formula for the standard error of savings during interval t of the reporting period:

$$\mathrm{Var}(\hat{s_t}) = \mathrm{var}(\widehat{kWh_t^P} - kWh_t^P)$$

$$= \mathrm{var}(a + b\, x_t^P - \alpha^P - \beta^P x_t^P - \varepsilon_t^P)$$

$$= \mathrm{Var}(a + b\, x_t^P) + \mathrm{Var}(\varepsilon_t^P)$$

$$= \widehat{\sigma^2} \mathbf{x_t^{P}}'(\mathbf{X'X})^{-1}\mathbf{x_t}^P + \sigma_P^2$$

where $\mathbf{x_t}^P$ is a 1 x 2 vector with first element equal to 1 and the second element equal to x_t^P. (Note: the two columns correspond to the two parameters of Equation 14 (α and β)). **X** is a T x 2 matrix, with ones in the first column and the values of x_t in the second column for the t=1, 2, …T intervals of the baseline period.

The third equality follows because α^P and β^P are unknown but fixed parameters, meaning their variance is zero and the error ε_t^P is independent. Note that the variance of the savings estimate for interval t depends on $\mathbf{x_t^{P}}'(\mathbf{X'X})^{-1}\mathbf{x_t}^P$—the variance of the *expected energy consumption* during

[40] For more detailed explanation of the OLS assumptions and unbiasedness theorem, see Sections 3.2 and 3.3 of Thiel (1971).

baseline conditions, conditional on x_t^P and on the variance of energy consumption during the reporting period $\hat{\sigma}_p^2$. The standard error is obtained by taking the square root of the variance.

Consistent with the definition of savings presented above, this derivation of the variance of forecast estimated savings assumes savings are estimated as a difference in expected energy consumption conditional on x_t^P. This implies that $\widehat{kWh_t^P}$ should be interpreted as the expected value of kWh, conditional on x_t^P under baseline conditions (i.e., $E[kWh_t| x_t^P, \alpha, \beta]$); kWh_t^P should be interpreted as the expected value kWh, conditional on x_t^P under SEM conditions plus an error. When taking the variance of kWh_t^P, it is necessary to account for the variance of ε_t^P.

The variance of the reporting period savings estimate $\widehat{S}$ can be determined through the variance of both sides of Equation 17:

$$\begin{aligned}
\mathrm{Var}(\widehat{S}) &= Var\left(\sum_{t=1}^{T^P} a + b\, x_t^P - \alpha^P - \beta^P x_t^P - \varepsilon_t^P\right) \\
&= \mathrm{Var}\left(\sum_{t=1}^{T^P} a + b\, x_t^P - \varepsilon_t^P\right) \\
&= \mathrm{Var}\left(\sum_{t=1}^{T^P} a + b\, x_t^P\right) + \mathrm{Var}\left(\sum_{t=1}^{T^P} \varepsilon_t^P\right) \\
&= \widehat{\sigma^2}\, \mathbf{x}^{Psum\prime}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^{Psum} + T^P \hat{\sigma}_p^2 \qquad \text{Equation 18}
\end{aligned}$$

where $\mathbf{x}^{Psum}$ is a 1 x 2 vector, with the first element equal to T^P and the second element equal to $\sum_{t=1}^{T^P} x_t^P$.

In Equation 18, making the simplifying assumption that the variance of the error in the baseline and reporting periods are equal (i.e., $\hat{\sigma}_p^2 = \widehat{\sigma^2}$), then the variance of reporting period savings equals[41]:

$$\begin{aligned}
\mathrm{Var}(\widehat{S}) &= \widehat{\sigma^2}\, \mathbf{x}^{Psum\prime}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^{Psum} + T^P\, \widehat{\sigma^2} \\
&= \widehat{\sigma^2}\left(\mathbf{x}^{Psum\prime}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^{Psum} + T^P\right) \qquad \text{Equation 19}
\end{aligned}$$

This derivation shows that the variance of the forecast savings estimate has two components: the first accounts for the variance of the estimated baseline model coefficients; and the second accounts for, in the reporting period, observing metered energy consumption (i.e., expected energy consumption conditional on x_t^P, plus an error) instead of expected energy consumption. Both components should be accounted for in estimating the variance of the savings estimate.

In addition to providing a more accurate estimate of the variance, accounting for the variance of metered energy consumption can help to explain unexpected results, such as an estimated increase in facility energy consumption intensity. For example, suppose that a facility experiences a random shock during the performance period that causes the facility's energy consumption to increase significantly and energy consumption intensity to increase. Since this

[41] Also, see Reddy and Claridge (2000), who derived a similar expression for the variance.

shock was large, it is important that the standard error of savings reflect the magnitude of the disturbance; otherwise, the standard error may be underestimated, the savings estimate may be reported as statistically significant (when it was not), and the evaluator may wrongly conclude that the program caused consumption to increase. Accounting for the error of metered energy consumption reduces the likelihood that the evaluator will find savings when none occurred.