Introduction

This document provides a set of model protocols for determining energy and demand savings that result from specific energy-efficiency measures or programs. The methods described here are approaches that are—or are among—the most commonly used in the energy-efficiency industry for certain measures or programs. As such, they draw from the existing body of research and best practices for energy-efficiency evaluation, measurement, and verification (EM&V).

These protocols were developed as part of the Uniform Methods Project (UMP), funded by the U.S. Department of Energy (DOE). The principal objective for the project was to establish easy-to-follow protocols based on commonly accepted methods for a core set of commonly deployed energy-efficiency measures.

1 About the Protocols

The methods described here represent generally accepted standard practices within the EM&V profession; however, they are not necessarily the *only* manner in which savings can be reliably determined. Still, program administrators and policy makers can adopt these methods with the assurance that: (1) they are consistent with commonly accepted practices, and (2) they have been vetted by technical experts in the field of energy 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 investments across programs, portfolios, and jurisdictions.

These protocols do not provide stipulated values for energy savings; however, the wide-spread use of these protocols would provide a common analytic foundation for determining "deemed" values while still allowing for the use of inputs appropriate for a project's particular circumstances. Nor do these protocols prescribe how baseline conditions should be defined for the calculation of impact for specific measures or programs (such as codes and standards or actual market conditions). Baseline conditions tend to vary with the type of measure or program, implementation method and whether the measures are for new construction, early-replacement, or replacement on burnout. Finally, these protocols do not prescribe specific criteria for either statistical confidence or the accuracy of savings estimates. Such thresholds are assumed to be set by the audiences, as determined by their unique objectives and priorities.

Instead, the protocols provide a structure for deciding on and applying such criteria consistently and for reporting the uncertainty associated with the indicated savings estimates.

2 Rationale

Investment in energy efficiency has increased steadily in the United States in recent years. In many jurisdictions, energy efficiency now accounts for a significant share of utilities' integrated resource portfolios. In several jurisdictions, energy efficiency has been 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 (EERS) 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, while in six jurisdictions, savings of over 20% are expected.[1]

With greater reliance on energy efficiency as a means of meeting future energy resource requirements, there is a growing demand for publicly available information on energy-efficiency programs, how their savings are measured, and how the achieved savings are reported. By the sharing and vetting of information among experienced practitioners and those new to the energy-efficiency field, this knowledge can reinforce the reliability of the savings. To this end, these protocols offer measure-specific evaluation methods and techniques for determining energy savings based on generally accepted practices in the energy-efficiency industry.

To help reduce the uncertainty associated with determining energy-efficiency savings, this material offers guidance for implementing the techniques and interpreting results. It can also provide a basis for comparing the impacts of energy-efficiency portfolios and policy initiatives across the country.

DOE envisions the following specific goals for this project:

- Offer guidelines that help strengthen the credibility of energy-efficiency program savings calculations.
- Provide clear, accessible, step-by-step protocols to determine savings for the most common energy-efficiency measures.
- Support consistency and transparency in how savings are calculated.
- Reduce the development and management costs of EM&V for energy-efficiency programs offered by public utility commissions, utilities, and program administrators.
- Allow for comparison of savings across similar efficiency programs and measures in different jurisdictions.
- Increase the acceptance of reported energy savings by financial and regulatory communities.

3 The Audiences and Objectives

In response to the interest of the State and Local Energy Efficiency Action Network (SEE Action)[2] EM&V Working Group and others, DOE commissioned this effort to provide a voluntary set of standard protocols for determining savings resulting from particular energy-efficiency measures and programs.

---

[1] See *Energy Efficiency Resource Standards: A Progress Report on State Experience*, American Council for an Energy Efficiency Economy (ACEEE), Report Number U112, June 2011.

[2] U.S. DOE: www.seeaction.energy.gov

While these protocols are applicable to a wide range of situations, their initial audience is expected to be stakeholders in states where energy efficiency is relatively new (or is newly expanded) *and* the issues of documenting savings have gained importance. From this general perspective, these protocols primarily serve evaluators working under the direction of regulators and/or program administrators in at least these four ways:

1. Providing a reliable basis for evaluating the effectiveness and viability of energy efficiency, thus offering regulators a basis and the means for both assessing the prudency of rate-payer-funded investments in energy efficiency and determining compliance with savings targets.
2. Offering utility resource planners and program administrators greater certainty about the performance of program performance and reduce planning and regulatory compliance risks.
3. Supplying independent EM&V contractors with a standard set of tools and techniques that would enhance the credibility of their findings.
4. Providing a resource for educating EM&V practitioners and a basis for the calculation of deemed savings in technical reference manuals (TRMs) that are being developed or updated in various jurisdictions.

By making the methods for calculation and verification of savings more transparent and uniform, these protocols will increase the reliability of energy-efficiency results reported by program administrators and implementation contractors. This will help mitigate the perceived risks of investing in energy efficiency and stimulate greater participation.

4 Definitions

Savings resulting from energy efficiency may be defined differently by various participants in the energy-efficiency industry (such as end-use energy consumers, project designers, contractors, program implementers and administrators, and utility resource planners—as well as independent, third-party evaluators). The UMP uses standard industry definitions to differentiate the four ways savings are reported at the design, implementation, and evaluation stages of a program's life cycle:[3]

- ***Projected Savings*** are values reported by a program implementer or administrator before the efficiency activities are completed.[4]
- ***Claimed (Gross) Savings*** are values reported by a program implementer or administrator after the efficiency activities have been completed. [5]

---

[3] Source: For more complete and detailed descriptions see the State and Local Energy Efficiency Action Network. 2012. *Energy-Efficiency Program Impact Evaluation Guide*. Prepared by Steven R. Schiller, Schiller Consulting, Inc. www.seeaction.energy.gov

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 net-to-gross (NTG) ratio.

- ***Evaluated (Gross) Savings*** are values reported by an independent, third-party evaluator after the efficiency activities and impact evaluation have been completed. The designations of "independent" and "third-party" are determined by those entities involved in the use of the evaluations and, thus, and may include evaluators retained by the program administrator or a regulator, for example.
- ***Net Savings*** are changes in energy use that are attributable to a particular energy-efficiency program. These changes may implicitly or explicitly include the effects of factors such as freeridership, participant and nonparticipant spillover, induced market effects, and changes in the level of energy service (e.g., rebound).

The UMP protocols provided here focus primarily on estimating evaluated gross first-year savings, except where estimates of net savings may be derived as part of the same method. The elements of net-to-gross (NTG) adjustments and the methods for measuring them will be described in a separate cross-cutting section dedicated to the topic in the second phase of this project. 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 is applied.

5 Project Process

The UMP project is a two-phase undertaking. This report, which presents the results of the first phase, contains protocols for these seven measures, which are primarily applicable to the residential and commercial facilities:

- Refrigerator recycling
- Commercial lighting
- Commercial lighting controls
- Residential lighting
- Residential furnaces and boilers
- Residential and small commercial unitary and split system air conditioning equipment
- Whole-building retrofit

These measures were selected because they: (1) represent a diverse set of end uses in the residential and commercial sectors; (2) are present in most energy-efficiency portfolios across all jurisdictions; and (3) have a significant remaining savings potential.

In the second phase, this list will be expanded, so the final set of measures covered is expected to represent a significant share of the available technical and economic energy-efficiency potential in most jurisdictions.

For each energy-efficiency measure, the protocol explains the underlying technology, the end uses affected by the measure, the method for calculating the measure's savings, and the data requirements. Also, each protocol attempts to provide a sufficient level of detail without being overly prescriptive, allowing flexibility and room for professional judgment.

The measure-specific protocols are supported and complemented by separate chapters that discuss technical issues and topics common to all measures. These cross-cutting topics, which are organized in the following five sections, are referenced in measure-specific protocols, where applicable:

1. Sample design
2. Survey design
3. Metering
4. Calculation of peak impacts, and
5. Other evaluation topics (including rebound and persistence of savings)

5.1 Relationship to Other Protocols

The protocols provided here are based on long-standing EM&V practices, and their methods conform to well-established engineering and statistical principles. They draw from and build on a number of previous attempts to develop comprehensive, systematic approaches to estimating the impacts of energy efficiency. Those efforts were conducted by various entities, including Oakridge National Laboratories (ORNL, 1991[6]), the Electric Power Research Institute (EPRI, 1991[7]), U.S. Environmental Protection Agency (EPA, 1995[8]), DOE, 1996[9] and DOE, 2008.[10]

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. In addition, a number of protocols have been developed to address specific EM&V requirements in certain jurisdictions (such as California and the Pacific Northwest).

A valuable companion document to this set of protocols is the *SEE Action Energy Efficiency Program Impact Evaluation Guide*. It 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.[11]

---

[6] Oak Ridge National Laboratory, Handbook of Evaluation of Utility DSM Programs, ORNL/CON-336, December 1991.

[7] Electric Power Research Institute. Impact Evaluation of Demand-Side Management Programs, Vol. 1: A Guide to Current Practice, EPRI CU-7179, Palo Alto, CA, February, 1991a.

[8] Conservation Verification Protocols, Version 2, EPA-430/B-95-012, June 1995.

[9] The North American Energy M&V Protocols, U.S. Department of Energy, DOE-GO 10096-248, February 1996.

[10] Federal Energy Management Program (FEMP) M&V Guidelines: Measurement and Verification for Federal Energy Projects Version 3.0, U.S. Department of Energy Federal Energy Management Program, April 2008.

[11] An initial examination of issues raised by pursuing a broadly applicable approach to EM&V can be found in National energy efficiency Evaluation, Measurement and Verification (EM&V) Standard:

Designed to be consistent 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.)

The EM&V methods described here also conform to the International Performance Measurement and Verification Protocol (IPMVP).[12] The UMP protocols expand upon the IPMVP options by adding detail and describing specific procedures for application to program- and portfolio-level evaluations. To this end, each protocol clearly identifies the IPMVP option with which it is associated.

6 Considering Resource Constraints

The UMP protocols are designed to represent approaches for providing accurate and reliable estimates of energy-efficiency savings that draw upon best practices without undue cost burdens. However, the UMP protocols do not offer recommendations regarding the levels of rigor and the specific criteria for accuracy of the savings estimates. Those issues are largely matters of policy, ease and costs of data acquisition, and availability of resources.

To provide maximum flexibility, each protocol contains 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. The costs of deploying the UMP protocols will vary, depending on the features of the energy-efficiency program being evaluated, the participant characteristics, and the desired levels of rigor and accuracy. 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.

7 About EM&V Resource Requirements

Historically, the costs of determining energy savings are embedded in the larger range of EM&V activities undertaken as part of large-scale programs for which public information is readily available. The range of those total evaluation costs can be obtained by reviewing those sources. For example,

- DOE's FEMP M&V Guidelines for federal-level performance contracting projects estimate the average, all-in cost of M&V as ranging from 3% to 5% of total project

---

Scoping Study of Issues and Implementation Requirements at http://www1.eere.energy.gov/seeaction/pdfs/emvstandard_scopingstudy.pdf

[12] International Performance Measurement and Verification Protocols, Concepts and Options for Determining Water and Energy Savings, Vol. 1, Prepared for the Energy Valuation Organization (EVO), January 2012.

costs.[13] The FEMP Guidelines report M&V expenses averaging 3.3% of costs for the typical performance-contracting project.[14]

- A report sponsored by NAESCO and the U.S. EPA suggests that each IPMVP Option will cost the client the following percentages of total project costs: from 1% to 5% for verification involving key parameters (IPMVP Option A), and from 3% to 10% for verification involving all parameters (IPMVP Option B).[15]
- In several jurisdictions, the evaluation costs for large demand-side management portfolios are available from regulatory filings. Our review revealed portfolio-level EM&V expenditures ranging from 3% of portfolio costs in Indiana to 4% of portfolio costs in California.[16]

As a general rule, 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 and research questions the EM&V activity aims to address.

7.1 Options for Small Utilities

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).[17] In these circumstances, program administrators may consider using deemed savings values from either:

- Technical reference manuals (TRMs) created by regional or state entities, or
- Values resulting from evaluations of similar programs performed by other regional utilities. (These can serve as the basis for determining energy-efficiency savings, provided that the installation and proper operation of the energy-efficiency measure or device has been verified.)

---

[13] FEMP M&V Guidelines, op. cit., p. 5-2.

[14] FEMP M&V Guidelines, op. cit., p. 5-9

[15] David Birr and Patricia Donahue, "*Meeting the Challenge – How Energy Performance contracting Can Help Schools Provide Comfortable, Healthy, and Productive Learning Environments*" (The National Association of Energy Services Companies and the US Environmental Protection Agency), pp. 32-33.

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

[17] According to the Small Business Administration, small utilities are currently defined as electric load serving entities with annual sales of less than 4 million MWh. Additional information on the costs and benefits of different measurement and verification approaches for small utilities can be found in the *Analysis of Proposed Department of Energy Evaluation, Measurement and Verification Protocols*, sponsored by the National Rural Electric Cooperative Association (NERCA) available at: http://www.nreca.coop/issues/ElectricIndustryIssues/Documents/EMVReportAugust2012.pdf

Deemed savings may be adjusted to allow for climate or other factors (regional or economic/demographic) that differ from utility to utility. Given the differences in how TRMs determine savings for identical measures, program administrators choosing this path should use deemed savings values that are 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).

Alternatively, 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 associations. (This has been done successfully in small utilities in California, Michigan, and the Pacific Northwest.)

Small utilities may also consider either coordinating with regional larger utilities or adopting the results of evaluations of similar programs implemented by larger utilities.

8 Project Management and Oversight

This project was funded by DOE and managed by the National Renewable Energy Laboratory (NREL). The Cadmus Group, Inc., was engaged to manage the protocol development and provide technical oversight. The project was designed to be inclusive of a broad set of stakeholders so as to ensure technical excellence. To facilitate the final appeal and acceptance of the work products, the following steps were taken.

8.1 Project Oversight by Variety of Stakeholders

NREL formed a project steering committee to provide general direction and guidance for the project. The steering committee consisted of regulators, utility managers, energy planners and policy makers, and representatives of industry associations.

8.2 Authorship by Experts

Nationally recognized experts on specific energy-efficiency measures and technologies drafted each protocol.

8.3 A Review by Technical Advisory Groups

Two four-member technical advisory groups—one focusing on the validity of the protocols and other on applicability—reviewed the drafts of the technical experts. These advisory groups consisted of experts from major consulting firms engaging in EM&V throughout North America.

8.4 A Review by Stakeholders

The protocols were subject to a review process that enabled stakeholders to provide feedback about the draft protocols before they were released in their final form.

Commercial and Industrial Lighting Evaluation Protocol

Dakers Gowans, Left Fork Energy

The Commercial and Industrial Lighting Evaluation Protocol (the protocol) describes methods to account for energy savings resulting from programmatic installation of efficient lighting 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. A separate Lighting Controls Evaluation Protocol addresses methods for evaluating savings resulting from lighting controls measures.

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, policy makers, 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]

1 Measure Description

An energy-efficiency measure is defined as a set of actions and equipment changes that result in reduced energy use—as 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, as 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 non-residential facilities include:

- Retrofitting existing, linear, fluorescent fixtures with efficacious[2] lamps and ballasts, or de-lamping over-lit spaces;
- Replacing incandescent lamps with compact fluorescent lamps (CFL); and
- Replacing high-bay fixtures (such as metal halide or linear fluorescent) with efficacious high-bay equipment (such as LED or high-performance linear fluorescent).

---

[1] As discussed under Section 7 of the Introduction chapter to the UMP Report, small utilities (as defined under the SBA 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 by "efficacy," with units of lumens per watt, where lumens are a measure of light output.

A separate Lighting Controls Evaluation Protocol addresses the methods accounting for savings resulting from lighting controls measures (such as adding time clocks, tuning energy management system commands, or adding occupancy sensors). In practice, lighting retrofit projects and new construction projects commonly implement lighting fixture and lighting controls measures concurrently, and this protocol accommodates these mixed measures

2 Application Conditions of the Protocol

Energy-efficiency lighting programs result in the installation of commercial, industrial, and non-residential 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. While the methods this protocol describes apply to all delivery mechanisms, issues with customer and baseline equipment data vary with each.

2.1 Common Program Types

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

2.1.1 Incentive and Rebate

Under this model, implementers pay program participants in target markets for installing lighting measures. Participants receive either an incentive payment, based on savings ($/kWh), or a rebate for each fixture or lamp ($/fixture, $/lamp). The terms incentive and rebate sometimes are used interchangeably but, generally, incentives are calculated based on project savings, while 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 and/or measured 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 trade-off: 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 facilitates determination of savings impacts.

Although rebate programs typically track useful information regarding 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. While residential equipment programs commonly use the upstream buy-down program delivery

approach, particularly for CFL lighting, commercial and industrial lighting programs use it less often.

Upstream buy-down programs do not interact with the end-use customers purchasing energy-efficient equipment; thus, baseline conditions and installation rates cannot be known. Program planners, implementers, and impact evaluators estimate these parameters, based on their experience with other programs or targeted market research studies.

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—overlooked by contractors working with incentive and rebate programs.

Direct-install programs can usually collect precise information about baseline and replacement equipment, so 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 acquire energy-efficiency reductions.

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 IECC 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 what would have occurred had the measure not been implemented (the baseline) and the energy consumption occurring after the retrofit. Energy calculations use the following, fundamental equation:

> Energy 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 can change between baseline and reporting periods.

Adjustments commonly apply to HVAC measures, but less commonly to lighting measures, or are inherent in algorithms for calculating savings.

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

> 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 1. Lighting Electric Energy Savings

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

where:

> kWh Save $_{light}$ = Annual kWh savings resulting from the lighting efficiency project
>
> fix watt $_{base,\ ee,\ i}$ = Fixture wattage, baseline or energy efficient, fixture type i
>
> qty $_{base,\ ee,\ i}$ = 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 $_{base,\ ee}$ = Annual hours of use, baseline or energy efficient, usually unchanged from baseline unless new controls are installed

Equation 2. Interactive Cooling Energy Savings for Interior Lighting

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

Equation 3. Interactive Heating Energy Savings for Interior Lighting

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

where:

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

kWh Save $_{\text{interact-heat}}$ = Interactive heating energy impact due to a lighting efficiency project, a negative value

IF $_{\text{kWh,c}}$ = Interactive cooling factor: the ratio of cooling energy reduction per unit of lighting energy reduction resulting from the reduction in lighting waste heat removed by an HVAC system

IF $_{\text{kWh,h}}$ = Interactive heating factor: the ratio of heating energy increase per unit of lighting energy resulting from reduction in lighting waste heat that must be supplied by an HVAC system during the heating season

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

Equation 4. Total Annual Energy Savings Due to Lighting Project

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

3.2 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 UMP document "Peak Demand and Time-Differentiated Energy Savings."

Equation 5. Lighting Electric Peak Demand Savings

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

where:

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

Equation 6. Interactive Electric Cooling Demand Savings for Interior Lighting

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

where:

kW Peak Save $_{\text{interact-cool}}$ = Interactive electric cooling demand impact due to a lighting efficiency project

$IF_{kWh,c}$ = 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 usually ignored in North America because heating equipment is typically non-electric and heating demand is usually not coincident with utility system peaks.

Equation 7. 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 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 (optional)
- Interactive factor for heating (optional)

Facilities—or spaces within facilities where the project is installed—are classified as cooled/uncooled or heated/unheated, so it is important to record information about heating and

cooling equipment and fuel types for each facility or space. This information is used to estimate interactive effects.

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; however, faxes, PDFs, and JPEGs do not meet these criteria.

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.

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 lookup table supplied by implementer, manufacturer cut sheet
Baseline fixture count	
Baseline fixture watt	From lookup table supplied by implementer, manufacturer cut sheet
Baseline HOU	From lookup table supplied by implementer; estimated by customer; BMS or meter data
Efficient fixture type	From lookup table supplied by implementer, manufacturer cut sheet
Efficient fixture count	
Efficient fixture watt	From lookup table supplied by implementer, manufacturer cut sheet
Efficient lighting HOU	Same as baseline if no controls installed
IF_c	Interactive factor for cooling, from lookup table, optional
IF_h	Interactive factor for heating, from lookup table, optional
kWh_{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.

5 Role of the Evaluator

The evaluator's role is to determine energy savings resulting from operation of lighting efficiency programs. The procedure is this:

1. Review a sample of completed projects, including conducting on-site M&V activities;
2. Calculate a realization rate (the ratio of evaluator-to-implementer anticipated savings); and
3. Use the realization rate to adjust the implementer-estimated savings.

5.1 Evaluator Data Requirements

The protocol recommends the program evaluator collect the same data as the implementer. As described in the Measurement and Verification Plan section, the evaluator must have access to the implementation lighting inventory forms and participant application material 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. 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_c	Interactive factor for cooling, from lookup table, optional
IF_h	Interactive factor for heating, from lookup table, optional
kWh_{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 projects. The M&V results are applied to the population of all completed projects to determine program savings. The sampling and application processes are described in the Sampling Design chapter.

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.[3]

[3] International Performance Measurement and Verification Protocol (IPMVP), which is considered the gold standard for evaluating energy-efficiency programs. IPMVP describes four options. Options A and B focus on

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.

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.

A description of the activities involved in the process follows this list of steps:

1. Select a representative sample of projects for review. (See the Sample Design chapter 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. (See Sample Design chapter for guidance on sampling.)

Confirm or correct the reported energy-efficient fixture type and wattage for each fixture in the sample.

Confirm or correct the reported quantity for all energy-efficient fixtures in the sample.

Confirm or correct the heating/cooling status and associated equipment for the spaces in the sample.

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 stored on-site for recycle or disposal.

4. Update lighting inventory form for the sample, based on findings from the on-site review.

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

---

retrofitted equipment, Option C computes savings at the facility or system level, often using utility billing records, and Option D uses computer simulation.

temporarily during the measurement period; however, some facilities have energy management systems that monitor lighting circuits, and these may be employed.

Metering equipment used to measure lighting operating hours either:

- Record a change of state (light on, light off), or
- Continuously sample and record 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

Establishing lighting operating hours is 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, 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 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 the lights on and off establish the threshold amperage for the on condition. 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 establish the light level threshold for the on condition.

While measuring amperage with data loggers is common, the continuous monitoring of light 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.

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 routines are programmed in an existing energy management system) during the verification site visit.

This process entails 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 is required per circuit; additional loggers may be deployed to offset logger failure or loss.
   B. 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 CT. Record the amperage threshold for the lights-on condition.

C. If using an energy management system, program trends for lighting on/off status for each circuit in the sample. The sampling interval should be 15 minutes or less. Check 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.

   A. For facilities with constant schedules (such as office buildings, grocery stores, and retail shops), the protocol requires metering for a minimum of two weeks.

   B. For facilities, 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 percentage of “on” time (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.

   A. For facilities with constant or variable schedules, the HOU parameter is calculated as: 8760 hours/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 period, times the percent-on time.

   C. The data used in the analysis should represent a typical schedule cycle, for example; 7, 14, 21 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 8760 hours/year, while classrooms are only active 6570 hours/year.

5. Evaluation timing requires the protocol meter operating hours after the efficiency project has been completed. The assumption in this process is that the operating hours have remained unchanged from the baseline period. Thus, HOU base and HOU ee in Equation 1 have the same value. (Note that will not be the case if the project includes lighting control measures.)

6. While a separate Lighting Controls Evaluation Protocol addresses lighting control measures, Equation 1 can accommodate changes in lighting operating hours, if measured data are available for the baseline period. This might occur in a facility with an energy management system or if a lighting contractor conducted a metering study before entering into a performance contract.

6.4 Report M&V 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 the measurement process and the sample lighting inventory form from the verification process, update the inventory HOU parameters, and calculate M&V savings for the sample.
2. Calculate the project realization rate: the ratio of M&V savings to the savings reported by the implementer for the project sample.
3. Calculate the project M&V savings, the product of the project realization rate, and the project savings reported by the implementer.

Project realization rates are averaged to create a *program* realization rate. The uncertainty and, therefore, the reliability of the program realization rate depend on the sample size (described later in the section "Sample Design"). 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.

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 the sections "Role of the Lighting Program Implementer" and "Role of the Evaluator," with additional detail in the M&V Plan section.

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 are updated as new fixtures and lighting technologies become available.

The wattage values are measured according to ANSI standards[4] by research facilities working on behalf of manufacturers and academic laboratories.

In the wattage table, each fixture and screw-in bulb are fully described and assigned a unique identifier. The *implementer* enters a fixture code into a lighting inventory form, which, if programmed, can be searched by a lookup function to show the associated demand. The

[4] 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% (see DOE 1993 Advanced Lighting Guidelines).

evaluator then verifies or corrects the fixture type for the evaluation sample, and updates the lighting wattage 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 provide examples (many others are available in most U.S. regions):

- Massachusetts Technical Reference Manual 2011, Massachusetts Device Codes and Rated Lighting System Wattage Table. Available from the Massachusetts Energy Efficiency Advisory Council, http://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, available from: http://www3.dps.ny.gov/W/PSCWeb.nsf/All/06F2FEE55575BD8A852576E4006F9AF7?OpenDocument. This is a comprehensive (34 pages) list, used by NYSERDA since the late 1990s, with recent data from California impact evaluation studies.
- Database for Energy Efficiency Resources (DEER). Available from the California Public Utilities Commission at: http://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 of a wattage table.

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 the "Measurement Process" section.

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 report project and program savings reliably. Accurate reporting by the implementer also results in more accurate evaluated savings for a given sample size.

The protocol requires program participants 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 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 provide more approximate estimates with 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 and Coincidence Factors

Energy-efficient lighting equipment produces less waste heat in building conditioned spaces, compared to the baseline. This results in a reduced cooling load and an increased heating load. Interactive factors—terms IF_c and IF_h in the "Algorithms" section—account for these additional changes in energy use.

Interactive effects are generally small for spaces conditioned for human comfort (2% to 6% for cooling in offices in New York City, for example[5].) They are also highly dependent on HVAC system types and efficiencies. For example, in a large office building in New York City, the IF_C varies with the equipment: (1) with gas heat and no economizer, the IF_C is 3.3%, (2) with an economizer, the IF_C is 1.9%, and (3) with economizer and VAV system, the IF_C is 6.5%.

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 factors are defined by space temperature ranges as follows[6]:

- Freezer spaces (-20 °F – 27 °F) = 50%

[5] TecMarket Works. October 2010. "New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs," Appendix D.

[6] Pennsylvania Public Utility Commission. 2011. "Technical Reference Manual." P. 138.

- Medium-temperature refrigerated spaces (28 °F – 40 °F) = 29%
- High-temperature refrigerated spaces (47 °F – 60 °F) = 18%
- Un-cooled space = 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 (electric or gas), the effect on other fuels is sometimes ignored. For example, electric energy-efficiency programs might report interactive electric cooling savings, but might omit interactive increases in gas heating energy.

Coincidence factors 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 to lower utility capacity requirements; thus, reducing the load on peak generation equipment that is usually the most costly to operate and improving 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.

Coincidence factors 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 coincidence factor of 1.0 for commercial lighting efficiency projects in New York State.[7] Typically, dawn-to-dusk exterior lighting has a coincidence factor of 0.0.

When accurate estimates of interactive values are available, the protocol recommends program implementers and evaluators use tables of interactive factors to report interactive effects for cooling and heating energy. The recommended sources for values of interactive factors, ranked by reliability, are these:

- Computer simulations of typical buildings found in the program's territory and weather zones;
- Interactive factors developed for similar programs and climates; and
- An average single value, developed from one or more tables of interactive factors for similar programs and climates.

Because the interactive effect is usually small relative to the primary energy savings from a lighting-efficiency project, program planners often borrow interactive factors developed for similar programs and climates.

The protocol also recommends using tables of coincidence factors (including any interactive effects from reduced cooling loads) to report system peak coincident electric demand savings. Like interactive factors, unique coincidence factors can either be developed for a program's territory and system peak characteristics or borrowed from programs with similar customer and utility profiles.

A sample of interactive and coincidence factors can be found in the documents cited in the reference section at the end of the protocol.

[7] TecMarket Works. P. 110.

7 Impact Evaluation

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

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, since difficulties and inaccuracies often occur when collecting data retroactively, particularly in attempts to back-fill missing data, determine baseline data, or deal with poor customer recall of project. 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 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.

7.1 Sample Design

The protocol requires sampling in two situations:

- To select projects from a program database for an impact study; and
- To select inventory lines for deploying light loggers.

Regulators normally prescribe the confidence and precision levels for the sample, or the implementer may impose them. (The separate Sample Design Protocol describes general sampling procedures.) The following details pertain specifically to lighting.

The protocol recommends stratified sampling when selecting projects for an impact study. 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%. (Refer to the Sample Design Protocol when developing a sampling plan.)

Light-logger studies also use stratified sampling for selecting samples, 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 (FEMP) M&V Guidelines[8] describe a detailed routine for selecting logging lines.

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

8 Other Evaluation Issues

8.1 Upstream Delivery

As upstream buy-down programs cannot access their individual customers, they lack the lighting inventory forms (with associated data) used to estimate savings. Implementers can use survey methods to estimate baseline fixture wattages and HOUs. Surveys require intercepting customers at the time of purchase to register their names and phone numbers.

Implementers 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.

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. Lighting power equals lighting power density (watts/ft2) times building area. HOUs are determined using the same methods as in incentive and rebate programs.

8.3 First Year vs. 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 usually will be less (sometimes dramatically so) than the product of first-year savings and measure life. The discount results from performance degradation and equipment failure or replacement. Lifetime savings are covered further in Other Evaluation Issues chapter; however, as lifetime savings for lighting projects are strongly driven by federal standards and changes in the market, they are discussed here.

Beginning in July 2012, most T12 lamps will not meet federal efficacy (lumens/watt) standards, accelerating a long-term trend towards T8 and T5 lamps and electronic ballasts. The effect is that first-year savings for T12 to T8 replacements can only be assumed 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; and

[8] http://www1.eere.energy.gov/femp/

- An efficient baseline with reduced savings for the remaining effective useful life.

The federal standard prohibits the manufacture of T12 lamps with current efficacy ratings. However, it is anticipated that sufficient stock will be available in the market for several years for burnout replacement. Regulators and administrators will need to consider T12 availability before instituting a dual baseline as a result of the standard.

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

8.4 Program Evaluation Elements

Building a 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 that the information required to conduct the research will be available.

9 References

The protocol depends heavily on reliable estimates of fixture wattages and HOU, CF, and IF values. A rich body of publicly available research exists to provide these data, which can be found in the references listed below. Although this is not an exhaustive list, it is representative. Users should select the reference best matching their market 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 but not all of these references provide source documentation.

Fixture wattages are generally based on manufacturer's 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.

California Energy Commission and California Public Utilities Commission (CPUC). 2008. "Database for Energy Efficient Resources (DEER)." http://www.deeresources.com.

Massachusetts Program Administrators. October 2011. "Massachusetts Technical Reference Manual for Estimating Savings from Energy Efficiency Measures 2012 Program Year – Plan Version." http://www.masssave.com/.

TecMarket Works. October 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://www3.dps.ny.gov/W/PSCWeb.nsf/All/06F2FEE55575BD8A852576E4006F9AF7?OpenDocument.

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

Pennsylvania Public Utility Commission. 2011. "Technical Reference Manual, Appendix C." http://www.puc.state.pa.us/electric/Act129/TRM.aspx.

Federal Energy Management Program. 2008. "M&V Guidelines: Measurement and Verification for Federal Energy Projects Version 3.0." http://www1.eere.energy.gov/femp/index.html.

California Energy Commission. 1993. "Advanced Lighting Guidelines."

10 APPENDIX

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**

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:.96-1.1)	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

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

New York Standard Approach for Estimating Energy Savings from Energy-Efficiency Programs 2010. Page 109.

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

Commercial and Industrial Lighting Controls Evaluation Protocol

Steve Carlson, DNV KEMA

The 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 these facilities:

- Commercial
- Industrial
- Government
- Institutional, and
- Other nonresidential facilities.

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 LEED) and the increased prevalence of demand-response programs.

This evaluation protocol, which focuses on lighting controls, is based on the assumption that the primary data captured will be used either to inform the evaluation or to determine deemed savings in a technical reference manual (TRM).

By following the methods presented here, evaluators can determine energy savings resulting from lighting controls installed through efficiency programs in a manner consistent across different jurisdictions or regions. The resulting data will provide planners, policy makers, regulators, and others with sound, comparable information for use in comprehensive energy planning.[1]

This protocol does not address savings resulting from changes in codes and standards, or from education and training activities.[2] 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 double counting of savings does not occur.[3]

---

[1] As discussed under Section 7 of the Introduction chapter to the UMP Report, small utilities (as defined under the SBA regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies, such as those presented here, should be considered for such utilities.

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

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

1 Measure Description

An "energy-efficient measure" can be defined as a set of actions and equipment changes resulting in reduced energy use—as compared to standard or existing practices—that maintain the same or improved service levels for customers or processes. Lighting controls are an energy-efficient measure because, in addition to delivering the light levels required for activities or processes in facilities, they can: (1) shut off lighting fixtures when a space is unoccupied, and/or (2) 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.[4] For new construction, the baseline conditions are generally provided by state and local building codes. However, these codes, which vary widely throughout the country, typically require some form of control for most interior lighting. This document contains 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);[5] and
- Continuous daylight dimming systems.

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).[6]

DRAFT

---

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

[5] Such HID fixtures typically have only one lamp that can be operated at two 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

This protocol applies to the installation in customer facilities of commercial, industrial, and nonresidential lighting control measures associated with energy-efficiency programs. While methods described in this protocol apply to all programs, the issues with customer and baseline equipment data vary with each.

Programs for commercial, industrial, and nonresidential lighting, which include lighting controls, can be classified as targeting either retrofit markets (serving existing facilities) or new construction markets. Various delivery methods are associated with these programs, and the method used depends on the target markets and customer types. The program delivery types described below apply to existing building programs.

The delivery method is typically distinguished by the party receiving incentive payments from a program. The most common program delivery types are: (1) incentive and rebate, and (2) direct install.

2.1 Incentive and Rebate

Under this delivery method, administrators pay program participants in target markets for installing lighting control measures. Participants receive an incentive payment that is based on either annual energy savings ($/kWh) for each installed measure or on demand savings ($/kW). Participants include design teams, contractors, building owners, or building managers.

The annual energy savings are estimated through one or both of these 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.

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

Incentive programs often collect more detailed baseline data than do rebate programs. This includes extensive data regarding controlled equipment wattages and hours of operation, which facilitates the 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 *ex ante* program savings.

2.2 Direct Install

Using this delivery method, contractors engaged through a program receive payment from that program for installing lighting control equipment in customer facilities. The customers receive a lighting control measure for free or at a reduced cost. Direct-install programs target hard-to-reach

DRAFT

customers—typically small businesses—who are often overlooked by contractors working through incentive and rebate programs. Direct-install programs typically do not focus on lighting controls, but these may be eligible measures.

2.3 New Construction Programs

New construction programs primarily employ incentives and rebates to achieve energy-efficiency reductions. However, these 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 IECC). Section 9.4.1 of the ASHRAE Standard defines lighting controls that are mandatory for interior lighting in buildings larger than 5,000 ft^2.[7] There may be federal, state, and local standards that contain additional baseline constraints on lighting controls.

3 Savings Calculations

The 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 this:

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

The equation's adjustment term calibrates baseline and/or reported use and demand to the same set of conditions. Adjustments are routinely made to account for changes in schedules, occupancy rates, weather, or other parameters that shift between baseline and reporting periods. While adjustments commonly are applied to HVAC measures, they are less commonly applied to lighting measures (or the 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

---

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

CSF = Control Savings Factor is the annualized reduction factor calculated across the EFLH

Equation 2A: Lighting Control Savings Factor

$$CSF = 1 - (EFLH_{post} / EFLH_{pre})$$

where:

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

$EFLH_{pre}$ = Annual equivalent full load hours before the application of controls

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

When calculating the site level CSF using measured data for multiple control points, develop the weighted average 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 due to 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; due to reductions in lighting waste heat removed by a HVAC system

$kWh_{interact –heat}$ = Interactive heating savings due to lighting control project (which is 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; due to reductions in lighting heat removed by a HVAC system

DRAFT

Equation 4: Total annual energy savings

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

3.1.1 Role of the Lighting Control Program Implementer

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

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

3.1.2 Implementation Data Requirements

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

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

Facilities (or spaces within facilities where the project has been installed) are classified as cooled/uncooled and heated/unheated. It is important to record information about heating and cooling equipment and the fuel types for each facility or space because this information is required to estimate interactive effects.

3.2 Implementation Data Collection Method

The protocol recommends participants collect and submit the required data as a condition for program enrollment. The protocol also recommends the implementer specify the formats for data reporting, 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 it must support combining multiple files into single tables for analysis by the implementer. Although 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 that are:

DRAFT

- Installed in an easily located room, floor, or space, and
- Belong to one usage group.

Table 1 lists fields required in the data reporting format.[8]

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
Control Savings Factor	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
kWhsave	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; and
- All lighting control equipment types and locations throughout the facility.
- For lighting control measures that reduce the power outputs of fixtures, describe the dimming controls, so each increment of light reduction has an appropriate kW value established.
- For daylight dimming systems, identify the locations and specify the minimum power levels of ambient light sensors. This will enable modeling of the system using building simulation software, if necessary. (Sensor location is not required when using a spreadsheet savings estimation approach.)

[8] The data sources for these fields are described in Section 5.4, Data Requirements and Sources, of this protocol

DRAFT

The protocol recommends integrating savings verification into the program administrative process, particularly regarding data tracking. Impact evaluations of lighting efficiency and lighting control 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.

To verify the accuracy of data provided on the lighting inventory forms, the administrator should employ a third-party expert to conduct periodic, systematic reviews and inspections of a sample of completed projects. Initially, the sampling procedure should be implemented randomly on a fixed-percentage basis (approximately 10%). This will ensure 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.

4 The Role of the Evaluator

The evaluator's role is to determine the energy savings resulting from lighting control efficiency programs. The procedure entails these steps:

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

4.1 Evaluator Data Requirements

This protocol recommends that both the program evaluator and the implementer collect the same data. As described in the Measurement and Verification Plan section, the evaluator must have access to the implementation lighting inventory forms and participant application materials for each project in the sample.

4.2 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 contains 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 data required for each project in the evaluation sample.

DRAFT

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

5 Measurement and Verification Plan

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

5.1 IPMVP Option

The selection of the appropriate International Performance Measurement and Verification Protocol (IPMVP)[9] evaluation method for reporting evaluated (*ex post*) savings is contingent upon site specific criteria. The key factors for determining the method are these:

- Availability of whole-premise, interval-metered data and, ideally, 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 use IPMVP Option C – Whole

[9] The IPMVP is considered the gold standard for evaluating energy-efficiency programs.

facility.[10] Option C requires that both pre- and post-metered data are available to evaluate the lighting control impacts. Since lighting controls often don't meet the relative impact criteria, IPMVP Option A (Retrofit Isolation: Key Parameter Measurement) approach is the most common method for evaluating *ex post* savings. The key parameters to be measured are $EFLH_{pre}$ and $EFLH_{post}$ for calculating the CSF term in Equation 1. Accurately measuring these variables typically requires determining lighting usage in the pre-control state (and measuring usage in the post-control state may also be required[11]).

Table 3 lists the metering requirements for measuring various types of lighting control measures. Table 3: Metering Requirements for Various Lighting Control Strategies

Lighting Control Measure	Metering Requirements		Metering Type
	Pre-Installation	Post-Installation	
Lighting Sweep Controls/EMS/Time Clock	Yes	Yes	Event or Power Logger
Occupancy Sensors	Yes	Yes/No	Event/Event &Occupancy Logger
Stepped Dimming (Dual Ballasts)	No	Yes	Event Logger
Dual Ballast (Hi/Low HID)	No	Yes	Power Logger
Continuous Daylight Dimming	No	Yes	Power Logger

In summary:

- Lighting sweep controls, EMS, and time clock measures require pre- and post-installation metering to establish EFLH and CSF accurately.
- Occupancy sensor measures can be determined effectively through pre-installation metering only if using a lighting event logger with infrared occupancy sensor capabilities.[12]
- Most dimming applications can be measured using post-installation data, but only when these are sufficiently accurate to support the assumption that uncontrolled kW would equal controlled lighting operating at full power.

DRAFT

---

10 In this case, the data could be either whole-premise data or lighting end-use data, which contains 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.

11 IPMVP Option A retrofit isolation requires the key savings variable be measured pre and post. However, when conducting M&V in an impact evaluation, obtaining baseline data can be challenging because the program administrator often does not collect the data, and the evaluators commonly do not become involved until after the project is installed.

12 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#

- Event loggers typically are lighting loggers that monitor lighting on/off operations via a photocell; power loggers monitor power consumption of controlled lighting circuits.

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

5.2 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. These are the steps for this process:

1. Select a representative sample of projects for review. (See the Sample Design chapter 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. This entails inspecting a representative sample of controlled lighting fixtures and lighting controls—reported by the implementer—to verify the controls are operating as reported. (See the Sample Design chapter for guidance on sampling.)
    a. Confirm or correct the 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, use physical evidence to supplement the interviews. (This evidence may be lighting controls installed on fixture types or in areas not changed by the project.)
4. Use the findings from the on-site review to update the lighting control inventory form for the sample.

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

5.3 Measurement Process

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

DRAFT

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

5.3.1 About Data Loggers

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

Change-of-state (event) lighting data loggers are small (matchbox-sized), integrated devices consisting of a photocell, a microprocessor, and memory storage. These data loggers are mounted inside of fixtures. Each time the lamps in a fixture 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, and lighting sweep controls/energy management systems. Event meters can also be used with stepped dimming systems, such as inboard/outboard configurations, where controlled lamps can be isolated from uncontrolled lamps.

For lighting control systems that offer varying lighting power (such as dimming systems or dual ballast HID systems, where the lamps cannot be isolated), the interval power of the lighting system must be monitored.

Data loggers that are continuously sampling and recording information about lighting operating hours 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 establish the 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 establishe the light level thresholds for "on" conditions. Data loggers are commonly used for amperage measurement; however, continuous light-level monitoring to determine hours of operation is less common.

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

5.3.2 About the Measurement Process

Measurement involves metering lighting operating hours for a representative sample of controlled fixtures selected for verification. Meters are deployed—or, if using an existing building management system BMS, metering routines are established—during the verification site visit. These are the activities for this process.

5.3.2.1 Meter operating hours for each circuit in the verification sample.

- 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.
- 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

DRAFT

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

- If the measure is an on/off type of lighting control (such as occupancy sensors) then use an event-type power logger. (These power loggers, which record a change of state when the power is on and off, provides similar data as change-of-state lighting loggers.) The sampling interval is irrelevant for an event logger since it captures transitions, so the data can be output at any interval desired.
- If using 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 BMS has sufficient capacity to archive recorded data. Also, confirm that the metering task will not adversely slow the BMS response time.

5.3.2.2 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 the deficiencies appear systemic, redeploy the loggers. If using BMS trends, spot-check the recorded data.

5.3.2.3 Leave metering equipment for the monitoring period.

The protocol recommends a monitoring period capturing the full range of facility operating schedules, so the metering period may include pre and post periods.

- 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 require additional time, which is based upon the variability of the schedules and the length of time necessary to capture the periodicity of the schedule changes.
- Facilities with seasonal schedules, such as schools, should be monitored during active periods.

5.3.2.4 Analyze metering data.

Calculate the percentage of “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 is necessary to annualize metered data to account for daylight hours during the metering period. This ensures that summer metering does not over-predict savings and winter metering does not under-predict savings. Similarly, facilities with seasonal schedules (such as schools, which should be metered during active periods) have annual EFLH and CSF values adjusted to reflect the operating schedules.

5.3.3 Report M&V Savings

Use the information collected during the M&V processes to calculate M&V project savings. These are the steps for this process:

DRAFT

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. Then calculate M&V savings for the sample.
2. Calculate the project realization rate (which is the ratio of M&V savings to savings reported by the implementer for the sample).
3. Calculate project M&V savings (which is the product of the project realization rate and the project savings reported by the implementer).
4. Use site-level savings estimates to develop program-level results. Site-level savings estimates are weighted and expanded upon based on the sample design to develop program-level realization rates and statistical relative precision at the selected confidence interval.[13]

5.4 Data Requirements and Sources

This section addresses information on the fixture wattage, EFLH, and CSF parameters in the algorithm equations. Data requirements are described in these sections: Role of the Lighting Controls Program Implementer and Role of the Evaluator. Additional details are provided in the M&V Plan section.

5.4.1 Fixture Wattage

Wattage values are measured according to American National Standards Institute (ANSI) standards[14] by research facilities working on behalf of manufacturers and academic laboratories. This protocol recommends using of fixture wattage tables that are developed and maintained by existing energy-efficiency programs and associated regulatory agencies. The tables, which list all common fixture types, are typically updated as new fixtures and lighting technologies become available.

Wattage tables are used by implementers and evaluators. 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, thus automatically updating lighting values.

---

[13] 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. However, ISO-NE requires relative precision of demand impacts be calculated at 80% confidence interval using a two-tail test. Both calculations provide the same result.

[14] 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 wattages vary, depending on the installation (suspended, recessed) and housing type. The differences end to be small—less than 5% (see *DOE 1993 Advanced Lighting Guidelines*.)

The protocol recommends adopting a fixture wattage table used by an established and recognized lighting-efficiency program.

5.4.1.1 Examples of Resources

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: http://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: http://www3.dps.ny.gov/W/PSCWeb.nsf/All/06F2FEE55575BD8A852576E4006F9AF7?OpenDocument. This is a comprehensive (34 page) list used by NYSERDA since the late 1990s.
- *Database for Energy Efficiency Resources.* Available from the California Public Utilities Commission: http://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.
- 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.

5.4.2 EFLH and CSF

EFLH and CSF greatly vary by application. To minimize uncertainty, the protocol requires *evaluators* do the following: (1) measure pre and/or post EFLH (depending on the control type, as listed in Table 3); and (2) calculate the CSF.

The following section describes data sources and methods used by *program implementers* to estimate EFLH and CSF parameters and reliably report project and program savings.

The protocol requires that program participants provide the following in their lighting inventory forms: (1) estimates of EFLH values by usage group, and (2) an estimate of CSF by control type. Note that the CSF estimate should *not* be based on the building schedule alone, although this may be used to inform the estimate.

The protocol recommends that participants—with guidance from the program implementer—develop EFLH and CSF values using one of the following sources:

- Lighting schedules in buildings with energy management systems or time clocks that control lighting equipment. To determine whether schedules are overridden, interview the building manager. When available, control schedules (or trend data) provide reliable estimates of true lighting operating hours.

DRAFT

- Interview with the building manager. Building managers are usually familiar with lighting schedules; however, they may not know how lighting is controlled, so they may not be a good source of estimates for CSF values.
- Tables containing EFLH and CSF values by building type (provided by the program implementer).
- Combinations of interviews and tables.

For calculating and reporting project savings, the protocol recommends requiring contractors for lighting-efficiency programs primarily use deemed EFLH-by-building-type values. Then, use 30% or less of the value for the CSF.

When EFLH values can be reliably estimated using site-specific data from an operating schedule organized by lighting control usage group, use these values 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 *when* the value does not exceed 50% of the published value.

Update deemed pre-control EFLH and CSF tables according to a continuous revision schedule, so that lighting programs using results from logger studies conducted for impact evaluation studies have current information. Table 4 lists the lighting control savings factors that were developed from ASHRAE 90.1 power adjustment factors.

Table 4: Lighting Control Savings Factors by Control Type

Lighting Control Type	Control Savings Factor
Light switch	0
No Controls	0
Daylight controls (DC) – continuous dimming	0.3
DC – multiple-step dimming	0.2
DC – ON/OFF	0.1
Occupancy sensor (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

DRAFT

6 Other Evaluation Issues

6.1 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. Overriding controls can also be utilized for retrofit and incentive programs, providing the site contact is cooperative, and the extra site visit is considered in

evaluation planning.[15] When possible, EFLH and CSF can be measured using pre/post metering techniques.

6.2 Coincidence Factor

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

Equation 3

$$kW_{reduction,\ coincident} = kW_{reduction} \times CF$$

where:

CF equals the percentage of the lighting load turned on during a utility peak period. (This varies with peak definition.)

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

7 Program Evaluation Elements

Building a foundation for the successful evaluation of a commercial, industrial, and nonresidential lighting controls program begins in the program design phase. Administrators must set data requirements before a program's launch, so when data are ultimately reviewed through an impact evaluation, the information required to conduct the research will be available.

Administrators support future evaluations by ensuring the data required for an impact study have been collected, stored, and checked for quality. These data include measured and stipulated values, which are available from prior studies or equipment tests.

DRAFT

[15] New Construction baselines may irrelevant, as lighting controls have mandatory provisions in recent standards (such as AHSRAE Standard 90.1-2004) requiring some form of lighting controls. For programmatic savings, any controls much exceed minimum baseline controls.

8 References

EFLH, CF, IF values and 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) *ANSI/ASHRAE/IESNA Standard 90.1-2004 Energy Standard for Buildings Except low-Rise Residential Buildings.*

California Energy Commission and California Public Utilities Commission. 2008. *Database for Energy Efficient Resources (DEER)*. http://www.energy.ca.gov/deer/

Massachusetts Program Administrators. 2011. *Massachusetts Technical Reference Manual for Estimating Savings from Energy Efficiency Measures 2012 Program Year – Plan Version.* http://www.masssave.com/

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. http://www3.dps.ny.gov/W/PSCWeb.nsf/All/

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

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

Federal Energy Management Program. 2008. *M&V Guidelines: Measurement and Verification for Federal Energy Projects Version 3.0.* http://www1.eere.energy.gov/femp/index.html

DRAFT

Small Commercial and Residential Unitary and Split System HVAC Cooling Equipment-Efficiency Upgrade Evaluation Protocol

David Jacobson, Jacobson Energy

Electricity savings from cooling equipment can be achieved by offering financial incentives to customers who install energy-efficient packaged (unitary) and split-system air conditioning equipment. This protocol applies to measures for residential and small commercial applications; however, it does not address early replacement incentive programs.

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 and 20 tons, although a system can be more than 100 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 rate using AHRI 340/360.

2 Measure Application and Delivery

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 is expressed as the energy-efficiency ratio (EER), seasonal energy-efficiency ratio (SEER), and integrated energy-efficiency ratio (IEER). The higher the EER or SEER, the more efficient the unit is.

- 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).
- SEER is a measure of a cooling system's efficiency over the entire cooling season for units under 65,000 BTU/hr (under 5.4 tons). The SEER, determined at part load, is measured at average conditions (82°F), as established by AHRI 210/240-2008.
- IEER is a measure of a cooling system's efficiency over the entire cooling season for units of 65,000 BTU/hr (5.4 tons) and above, expressed in Btu/hr of cooling per watt of electric input. AHRI Standard 340/360 2007 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

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

of at various load capacities." It replaces the Integrated Part Load Performance (IPLV) in AHSRAE standard 90.1-2007.

For many commercial unitary rebate programs offered in 2011 and 2012, units greater than 5.4 tons are qualified based on the EER only, and IEER is not captured. Although IEER provides a more accurate measure of seasonal efficiency for larger units, it is not yet commonplace throughout the incentive program community.

Table 1 presents a typical program offering for this measure.[2]

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

Unit Size			Efficiency Tier			
			Level 1		Level 2	
Tons	**Btuh**		**Min. SEER/EER for Incentive**	**Incentive $/Ton**	**Min. SEER/EER for Incentive**	**Incentive $/Ton**
< 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	$80
≥ 11.25 to < 20	≥ 135,000 to < 240,000		11.5 EER	$50	12.0 EER	$80
≥ 20 to < 63	≥ 240,000 to < 760,000		10.5 EER	$30	10.8 EER	$50
≥ 63	≥ 760,000		N/A	N/A	10.2 EER	$50

As noted, this measure's primary delivery channel is a rebate program, usually marketed through program administrator staffs and heating ventilation and air conditioning (HVAC) contractor partners. Typically, these programs do not include early replacement incentives, except when unusually high use of air conditioning occurs.

- Rebates for units installed in commercial settings are typically paid on the basis of dollars-per-ton of cooling, which can vary by the efficiency level achieved (CEE 2009).

- Rebates for residential units are often paid on a fixed rebate-per-unit basis to discourage oversizing and to promote high-quality installation practices.

The rebates apply either: (1) at the time of normal replacement due to age or failure, or (2) for new construction applications.

When a unit is installed in new construction or replaces an existing unit that has failed, the baseline efficiency standard it must meet is generally defined by local energy codes, federal manufacturing standards, or ASHRAE Standard 90.1 for SEER-rated units (below 5.4 tons) and IEER-rated units (5.4 tons or greater). This protocol assumes more-efficient equipment of the same capacity runs the same number of hours as the baseline equipment. It does not cover:

[2] MassSave Cool Choice Program, offered in 2012 by all Massachusetts Program administrators. See http://www.masssave.com/~/media/Files/Professional/Applications-and-Rebate-Forms/Cool_Choice_MA_Form_fnl.ashx

- Early replacement retrofits
- Right-sizing initiatives
- Tune-ups
- ECM motor retrofits
- Savings resulting from installation of an economizer or demand-controlled 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 electronically commutated motors (ECM) for ventilation fans as a retrofit or as 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 (ACCA) Quality Installation (QI) standards, which encompasses proper duct sealing (ACCA 2007).

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 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 newer packaged or split systems have temperature or dry bulb-based economizers, as required by code or by standard practice. Units with temperature-based economizers can be included in samples as a random occurrence, reflected in approximately rough 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 are not required by code, some utilities provide an incentive for them. If programs offer additional incentives for dual-enthalpy 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. However, units that receive rebates for demand-controlled ventilation should not use this protocol, which assumes the operating hours remain constant.

DRAFT

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 installed 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, as defined by a large number of technical reference manuals[3] (TRMs), uses the following algorithms.

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

$$\text{kWh Saved} = (\text{Size kBtu/hr}) \times (1/EER_{baseline} - 1/EER_{installed}) \times (EFLH)$$

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

$$\text{kWh Saved} = (\text{Size kBtu/hr}) \times (1/SEER_{baseline} - 1/SEER_{installed}) \times (EFLH)$$

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	= Equivalent full-load hours for cooling

Although at this time, many efficiency providers use Equation 2 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)

$$\text{kWh Saved} = (\text{Size kBtu/hr}) \times (1/IEER_{baseline} - 1/IEER_{installed}) \times (EFLH)$$

Where:

[3] Massachusetts Program Administrators (2011); United Illuminating Company and Connecticut Lighting and Power Company (2008); Vermont Energy Investment Corporation (2010).

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

$IEER_{installed}$ = Seasonal energy-efficiency ratio of the specific high-efficiency unit

Note that for many programs currently offered, only EER is required to qualify units of greater than 5.4 tons. For smaller units, SEER is almost always available, and it should be used for the calculation of annual energy savings.

This formula assumes 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 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; and
- 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, utilize these primary components:

- Unit size
- Efficiency of the base unit and the installed unit
- Annual operating hours for energy savings
- Coincidence factor (CF) 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 IPMVP[4] 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

[4] International Performance Measurement and Verification Protocol (IPMVP), which is considered the gold standard for evaluating energy-efficiency programs.

cannot be measured or assessed for participating customers because, by definition, lower-efficiency baseline equipment was never installed. (That is, the unit replaced is often old and below current requirements.) 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.[5]

- Regarding the use of pre/post-billing analysis (IMPVP 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 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 required. As more residential and small commercial customers get kW interval data, using post-installation data to get overall facility cooling hours is more viable

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, on average, units perform close 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 equation listed in Section 3, 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 models can be developed for a wide variety of building types, system configurations, and applicable weather data. Such analysis usually results in an extensive

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

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; and 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 program planning estimates when measurements are not available, 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 energy consumption (kW/kWh) for a sample of units to develop EFLH estimates (KEMA 2010). Note that the energy consumption of the compressor, condenser, evaporator, and supply fans are used to calculate the EFLH, but only when the compressor and condenser actually supply cooling. Measurement of 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 kWh 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 kW/kWh usage 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 kWh consumption by the peak rated kW consumption serves as a proxy for EFLH. The connected load is defined as a unit's peak cooling capacity at AHRI conditions in kBTU/hr and divided by the EER. Such metering should be true power kW metering, which, at a minimum, involves the compressor and condenser fan. Ideally, however, all components would be metered, including the supply fan and evaporator fan. If kW metering proves too costly, the amperage data may be acceptable if it is supplemented with spot wattage measurements under a variety of loading conditions.

When taking measurements, consider these factors: (1) Use 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.)

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 Secondary Options

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.

DRAFT

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°F used for EER ratings and the humidity is very low, compared to conditions for SEER ratings. Navigant[6] has shown that performance in hot, dry climates differs significantly from manufacturers' standard conditions.

Another complicating issue is performance at low loading for large units, with multiple compressors running in parallel. 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.

Use manufacturers' detailed performance data for analysis of unitary and split-system equipment where cooling is a very large part of a portfolio ***or*** where part-load operation is critical to unit performance ***and*** typical operating conditions are far from IEER or SEER conditions. The basic method is as follows:

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 and EER or SEER and EER.
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 involvedmeasurement 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 AHSRAE publications.[7])

---

6 Navigant, 2010

7 ASHRAE 2000, 2002, 2010

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.[8]

4.3 Verification Process

The key data to be verified are these: (1) the size of the unit rebated; and (2) the 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) or
- An on-site audit of a sample of participants (usually the same participants selected for the end-use metering, discussed above).

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.

4.4 Data Requirements

Minimum data required for evaluating a unitary HVAC rebate program are these:

- Size (in BTU/hr or tons) of each unit installed;
- Efficiency (in EER, SEER, or IEER) of each unit installed;
- Assumed baseline efficiency for each category of units (from prevailing code or standard); and
- 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 the Section 3, Savings Calculations, determine the savings for this measure with these data:

- The installed cooling capacity

[8] As discussed under Section 7 of the Introduction chapter to the UMP Report, small utilities (as defined under the SBA regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

DRAFT

- The EER, SEER, or IEER rating (from manufacturers' data) of the baseline unit and the installed unit,
- The measured EFLH.

4.5 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; and KEMA, 2010). Further, at least two studies have examined full-load hours of residential central air conditioning systems (KEMA, 2009; and ADM, 2008). The method this protocol recommends has been based on work described in the Northeast Energy Efficiency Partnerships (NEEP) EM&V Forum study, 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. Experience has generally shown that large units tend to run for more hours and exhibit higher peak coincidence than small units (ranging from 3 tons to 20 tons). Large units also tend to use multiple compressors, which are controlled differently than smaller, single-compressor units.

If a program predominantly rebates units less 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.

4.5.1 Metering

Metering should involve taking true RMS kW power measurements at one-minute intervals during at least half of the warm weather period and either the spring or fall shoulder periods. Preferably, metering should extend from the time units typically come on in spring until units are no longer needed in fall.

As recommended in this protocol, the one-minute intervals allow data analysis of cycling patterns beyond the determination of EFLH. Data will be aggregated to one-hour averages for use in the model specified below.

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

DRAFT

supplying cooling. The accuracy of kW measurements should be ± 2%, as recommended by ISO 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 8760 in kW
THI_{dh}	= Temperature-humidity index on day *d* hour *h*
w(d)	= 0/1 dummy indicating day type of day *d*, Monday = 1, Sunday =7, Holiday = 8
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
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}	= the outside dry bulb temperature in °F, and
DPT	= the outside air dew point temperature in °F

Note that, that this particular functional form is just an example of what has been successfully used. 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

[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

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 root mean squared error (RMSE), R-square for the model, and the mean bias error.

The following equation provides an EFLH calculation for the overall load shape 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.

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.

The EFLH for the population can be determined by multiplying the EFLH for each metered unit by the appropriate weighting factor, 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 either: (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.

4.6 Sample Design

Evaluators will determine the required targets for the confidence and precision levels, subject to specific regulatory or program administrator requirements. 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 group.

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.

DRAFT

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. CVs generally range from 0.5 to 1.0[10], and the more homogeneous the population, the lower the likely CV. After the study is completed, recalculate the CV and determine the estimate of sampling error.

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. Also, 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.

5 Program Evaluation Elements

To assure 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.).
- When units to be metered are 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.)
- When units to be metered malfunction during the mid-metering period and have (or have not) been repaired at the customer's instigation.

[10] At a CV of 0.5, the sample size to achieve 90/10 is 67. At CV of 1.0, the sample size is 270.

It is recommended to add to the sample an additional 10% of the number of sites or units to account for data attrition.[11]

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.[12]

One key issue is how to extrapolate data beyond the measurement period for 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 turned off for the season. These interview data can be used to override regression analysis indicating usage 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° F, most units shut off for the season. That information enabled KEMA to apply this rule to other sites in the NEEP EMV Forum study, resulting in a more realistic estimate of fall and winter cooling hours than was obtained by applying only regression results.

5.1 Net-to-Gross

A separate cross-cutting protocol to determine applicable net-to-gross is currently being prepared.

DRAFT

[11] In KEMA's study for the NEEP EMV Forum, approximately 9% of metered units were removed due to data validity problems (KEMA, 2011).

[12] The Residential Lighting Protocol, further discussed in-service rates.

6. Bibliography

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

American Society of Heating Refrigeration and Air Conditioning Engineers , (2002), "ASHRAE Guideline 14-2002 Measurement of Energy and Demand Savings"(Revision 14-2002R in process).

American Society of Heating Refrigeration and Air Conditioning Engineers (2010),, "Performance Measurement Protocols for Commercial Buildings",

ACCA. (2007). Air Conditioning Contractors of America (ACCA) Standard 5 (ANSI/ACCA 5 QI-207) HVAC Quality Installation Specification.

ADM (December 2008). "Residential Central AC Regional Evaluation." Prepared for NSTAR Electric and Gas Corporation, National Grid, Connecticut Light & Power, and United Illuminating.

Air-Conditioning, Heating and Refrigeration Institute (AHRI). (2008). "ANSI/AHRI 210/240-2008 with Addendum 1, Performance Rating of Unitary Air-Conditioning & Air-Source Heat Pump Equipment."

Consortium for Energy Efficiency (CEE) Commercial Unitary AC and HP Specifications, Unitary Air Conditioning Specification, Effective January 16, 2009, http://www.cee1.org/com/hecac/hecac-tiers.pdf.

Consortium for Energy Efficiency (CEE). (January 2010.) "Information for CEE Program Administrators on the New Part Load Efficiency Metric for Unitary Commercial HVAC Equipment." http://www.cee1.org/com/hecac/Prog_Guidance_IEER.pdf

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.

KEMA. (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."

DRAFT

KEMA. (August 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).

KEMA. (March 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.

Massachusetts Program Administrators. (October 2011). “Massachusetts Technical Reference Manual for Estimating Savings from Energy Efficiency Measures 2012 Program Year – Plan Version.”

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.

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 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.

The 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 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.

DRAFT

Residential Furnaces and Boilers Evaluation Protocol

David Jacobson, Jacobson Energy

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.

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) 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

[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] CenterPoint Energy's high-efficiency heating system rebate program offered in 2011. See: http://www.centerpointenergy.com/services/naturalgas/residential/efficiencyrebatesandprograms/heatingsystemrebates/MN/

[3] MassSave/GasNetworks 2012 High Efficiency Heating and Water Heating Rebates for Residential Customers. See: http://www.masssave.com/~/media/Files/Residential/Applications%20and%20Rebate%20Forms/2012%20GN%20Rebate.ashx

DRAFT

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;
- Replacing equipment at the end of its normal life or upon failure; or
- 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 another 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 electrically 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 determine 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 these:

- What data are collected at the time of installation or application for incentives?
- What data can be easily collected during an evaluation?

DRAFT

- 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 formulae 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 capacities (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/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 provision of capacity information—and how this affects derivation of formulae for calculating savings—the recommended formula presented in the two forms depends on the capacity value provided. (Appendix A provides derivations for calculating savings through these two methods.)

- The first derivation assumes data collected regarding unit size is input heating capacity.
- The second derivation assumes size data collected is output capacity.

[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.

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

$EFLH_{e\text{-}installed}$ = full-load equivalent 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

$$Savings_{b-e} = Capacity_{output} * ELFH * (1/ AFUE_b - 1/ AFUE_e)$$

Where:

$Capacity_{output}$ = heating output capacity of both the baseline and installed high-efficiency unit

ELFH = 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

---

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 (i.e., the installer, knowing the unit is more efficient—or relying on the ratings—will install a unit with less input requirements for the higher-efficiency unit). This makes engineering sense but, again, it depends on whether input or output ratings are used.

$AFUE_e$ = Annual Fuel Utilization Efficiency of the high-efficiency unit

Note that the $Capacity_{output}$ * ELFH equals the annual heating (BTUs 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; and
- The cost, complexity, and uncertainty in measuring each of those variables.

4.1 IPMVP[8] 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, KEMA [2009]; and 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 sub-metering, it is not feasible to conduct direct gas sub-metering of a sufficiently large sample to represent varying types of equipment (boilers and furnaces with varying efficiency levels) and of home and homeowner

[8] International Performance Measurement and Verification Protocol (IPMVP), which is considered the gold standard for evaluating energy-efficiency programs.

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 serves as the recommended 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; and
- The availability and relevance of billing data.

The primary variables for determining savings for high-efficiency boiler and furnaces are these:

1. The installed unit size or capacity in Btu/hr (either input or output).
2. The AFUE efficiency rating of baseline unit.
3. The AFUE efficiency 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 this: 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 two factors: (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 size (Btu/hr) and the 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 these:

- 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. While 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,

[9] Often listed as GAMA in the manufacturer's literature.

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.

5 Discussion of Methodology

The methodology used to calculate savings for each unit and, if required, to calculate the corresponding ELFH, 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

$\text{EFLH}_{\text{e-installed}}$ = full-load equivalent hours of the installed high-efficiency unit

Assuming the Gas Used for Heating = Normalized Annual Heating Consumption (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,

[10] http://www.ahridirectory.org/ahridirectory/pages/home.aspx

combined with sufficient post-installation billing data, allows calculation of savings using a billing analysis for a high percentage of participants.

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). The 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): 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 the 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 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 CCF per heating degree day

H_0 = the average normal heating degree days

α. = non-heating usages in therms of CCF[11] 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.

[11] 1 CCF = 100 cubic feet of natural gas.

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 ECM motors 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; and
- Steam boilers over 150,000 Btu/hr.

If an evaluation seeks to update variables in a TRM, either Equations 1 or Equation 2:

Equation 1

$$\text{Savings} = \text{Capacity}_{\text{input-e}} \times \text{ELFH}_e \times ((\text{AFUE}_e / \text{AFUE}_b) - 1)$$

or

Equation 2

$$\text{Savings} = \text{Capacity}_{\text{output-e}} \times \text{ELFH}_e \times (1/\text{AFUE}_b - 1/\text{AFUE}_e)$$

In each case, EFLH_e can be determined by Equation 7:

Equation 7

$$\text{EFLH}_e = \text{NAH}_e / \text{Capacity}_{\text{input-e}}.$$

The equation used is determined by:

- Whether the program collects $\text{Capacity}_{\text{input}}$, $\text{Capacity}_{\text{output}}$, or both as part of the application and data collection process; and
- 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 these:

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.

DRAFT

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 either: (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 than 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 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.[12] The Pooled Fixed-Effects Approach Section 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

$$\Delta NAH_k = \gamma_{Hk} H_{0k} + \Sigma_q \gamma_{Hkq} H_{0k} x_{qk}$$

[12] Again, the approach recommended for this process is discussed in the Whole Building Retrofit Protocol.

where the data, model structure, and estimation procedures are as described in the Whole Buildings Retrofit Chapter.

The two-stage, site-level modeling approach discussed in the Whole Buildings Retrofit Chapter 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

$$Savings_{e-replaced} = \Delta NAH_{e\text{-}replaced} = NAH_e*((AFUE_e / AFUE_{replaced}) - 1)$$

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, or examples of specific models, manufacturers, and capacities.[13]

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, which would underestimate savings. 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.

[13] Preston's guides provide a good resource for efficiency specifications on old units: http://www.prestonguide.com/

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 ELFH.

If average of the $Capacity_{input}$ is known for each category, then:

Equation 13

$$ELFH = Savings_{e-b} / (Capacity_{input} * ((AFUE_e / AFUE_b) - 1))$$

or

Equation 7

$$EFLH = NAH_e / Capacity_{input\text{-}e}.$$

As discussed in the Introduction chapter to the UMP Report, small utilities—as defined under the SBA regulations—may face additional hurdles and constraints in undertaking this Residential Furnaces and Boilers Evaluation Protocol. Therefore, alternative methodologies, (also discussed under Section 7), should be considered for such utilities.

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.

DRAFT

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 error 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 these:

- 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; and
- The procedures for filling in limited amounts of missing billing data.

6.2 Net-to Gross

As the methods discussed above do not involve a control group, it is necessary to determine an appropriate net-to-gross ratio. For specifics on the recommended method for determining a survey-based estimate of self-reported freeridership and spillover from participants and contractors, see the Net to Gross chapter.

DRAFT

7 References

Fels, Margaret, F. 1986. "PRISM: An Introduction." Energy and Buildings, 9, 5- 18.

KEMA. 2009. *New Jersey's Clean Energy Program Residential HVAC Impact Evaluation and Protocol Review*.

KEMA. 2008. *Puget Sound Energy's Residential Energy Efficient Furnace Program Impact Evaluation*.

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. 2011. *Protocols to Measure Resource Savings*.

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

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.

Vermont Energy Investment Corporation. 2010. *State of Ohio Energy Efficiency Technical Reference Manual*.

DRAFT

7. Appendix A. Simplified Formulas for Calculating Savings from Upgrading the Efficiency of a Residential Gas Furnace or Boiler

7.1 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:

$$Q_{loss} = Capacity_{input} * EFLH_{\mathbf{b}} * AFUE_{\mathbf{b}}$$

$$Q_{loss} = Capacity_{input} * EFLHe * AFUE_{\mathbf{e}}$$

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:

$$Savings = Capacity_{input} * EFLH_{\mathbf{b}} - Capacity_{input} \times ELFH_{\mathbf{e}}$$

$$= Capacity_{input} * EFLH_{\mathbf{e}} * (AFUE_{\mathbf{e}} / AFUE_{b}) - Capacity_{input} \times ELFH_{\mathbf{e}}$$

$$= Capacity_{input} * EFLHe * ((AFUE_{\mathbf{e}} / AFUE_{b}) - 1)$$

To use the Normalized Annual Heating consumption (NAH) of gas for heating from billing data, via a degree day-based regression analysis or end-use metering, apply this equation:

$$NAH_{\mathbf{e}} = Capacity_{input} * EFLH_{\mathbf{e}}$$

Substituting the above into the savings equation produces:

$$Savings = NAH_{\mathbf{e}} * ((AFUE_{\mathbf{e}} / AFUE_{\mathbf{b}}) - 1)$$

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:

$$Savings = Capacity_{input} * EFLHe * (AFUE_{\mathbf{e}} / AFUE_b) - 1)$$

7.2 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 = $Capacity_{output} / Capacity_{input}$

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 ($Capacity_{output\text{-}e}$) of unit and hours are the same for each unit:

$$Annual\ Heating\ Load = Capacity_{output\text{-}e} * ELFH$$

Then:

$Savings = Input\ Energy_b - Input\ Energy_e$ = Annual Heat Load / $AFUE_b$ – Annual Heat Load / $AFUE_e$

= Annual Heat Load * (1/ $AFUE_b$ – 1/ $AFUE_e$)

= $Capacity_{output}$ * ELFH * (1/ $AFUE_b$ – 1/ $AFUE_e$)

Rearranging:

$$Savings = Capacity_{output} * EFLH / AFUE_e * ((AFUE_{\mathbf{e}} / AFUE_{\mathbf{b}}) - 1)$$

Noting that:

$Capacity_{output} / AFUE_e = Capacity_{input\text{-}e}$ and $NAH_{\mathbf{e}} = Capacity_{input\text{-}e} * EFLH_{\mathbf{e}}$

Yields the same equations as above:

$$\text{Savings} = NAH_{\mathbf{e}} * ((AFUE_{\mathbf{e}} / AFUE_{\mathbf{b}}) - 1)$$

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:

$$\text{Savings} = Capacity_{output} * ELFH_e * (1/ AFUE_b - 1/ AFUE_e)$$

DRAFT

Residential Lighting Evaluation Protocol

Scott Dimetrosky, Apex Analytics, LLC

This Residential Lighting Evaluation Protocol addresses evaluation issues for all variants of residential lighting energy-efficiency measures, and it highlights differences in approaches resulting from variations in program design and delivery mechanisms.

1 Residential Lighting Measure Description

In recent years, residential lighting has represented a significant share of ratepayer-funded electricity energy-efficiency savings. The majority of these savings have been achieved by promoting the purchase and installation of compact fluorescent lamps (CFLs), both standard "twister" bulbs and specialty CFLs such as reflectors, A-Lamps, globes, and dimmable lights. Some efficiency programs have also promoted ENERGY STAR® lighting fixtures. More recently, programs are introducing solid-state light emitting diode (LED) lamps.

The future of savings claims from residential lighting programs is uncertain, due to the provisions of the 2007 Energy Independence and Security Act (EISA). This legislation requires that most screw-based light bulbs become approximately 28% more energy efficient during the period from 2012 through 2014, as measured by the efficacy in units of lumens per watt (W). EISA requirements take effect in phases, beginning with 100-W equivalents in 2012, 75-W equivalents in 2013, and 60- and 40-W equivalents in 2014. To add further uncertainty regarding the baseline, the federal spending bill approved in December 2012 eliminated enforcement of the EISA standards through at least September 2012.

2 Application Conditions of Protocol

Residential lighting measures are typically delivered by program administrators through four mechanisms:

- ***Upstream Buy-Down/Mark-Down.*** The most common approach to achieve residential lighting savings has been through "upstream" incentives to either manufacturers to buy down (or have retailers mark down) the cost of lights for consumers. This delivery mechanism offers the discount at the time of purchase (that is, at the point of sale) and, thus, does not require any application or paperwork from the end-use customer.
- ***Direct Installation.*** Many program administrators who offer residential audit programs also provide direct installation of CFLs at the time of the audit. In most programs, the audit is offered at either no cost or at a highly discounted cost to the customer, and there is usually no additional cost for the CFLs.
- ***Giveaways.*** A number of program administrators have provided CFLs free of charge to residential customers through the mail, at customer service offices, or at community, religious, or local government events. In some programs, the CFLs are mailed to customers only upon request. In other programs, the CFLs are distributed without prior customer request. The amount of customer information collected at the

DRAFT

time of giveaway events varies, with some program administrators requiring full name and contact information and other program administrators not requiring any.

- ***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 fill out their name and contact information to obtain the product at the discounted price or to receive the rebate.

While 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. Where necessary, this protocol highlights and provides more detail regarding specific differences. Also, program administrators may need to prioritize their evaluation resources on particular 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 Savings Calculations[1]

Gross energy first-year savings from residential lighting measures can be calculated through a number of different algorithms. The approach recommended is based on the following general algorithm:

Equation 1

$$kWh_{saved} = NUMMEAS * (\Delta W/1{,}000) * HRS * ISR * INTEF$$

kWh_{saved}	First-year electricity savings measured in kilowatt-hours
NUMMEAS	Number of measures sold or distributed through the program
ΔW	Delta watts = baseline wattage minus efficient lighting product wattage
HRS	Annual operating hours
ISR	In-service rate
INTEF	Cooling and heating interactive effects

The recommended techniques for estimating each of these parameters, based on either primary or secondary data, are described in this chapter.

[1] 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.

DRAFT

4 Measurement and Verification Plan

The savings from residential lighting measures should be calculated through a mix of measured and estimated parameters. This approach, which is similar to Option A of the International Performance Measurement and Verification Protocols[2] (IPMVP), is recommended because the values for some parameters can be directly measured through metering (such as annual hours-of-use), while others parameters (such as delta watts for upstream lighting programs) need to be estimated through other techniques.

4.1 Number of Measures Sold or Distributed

The number of measures sold or distributed through a program should be collected by the administrator or a third-party implementation contractor. Data should be compiled in electronic format in a database that tracks as much detail as possible regarding the measures delivered. For example, for an upstream program, this should include detailed information for each transaction:

- Product shipment dates from manufacturer to retailer, where applicable
- Detailed product information
  - Bulb type (e.g., CFL, LED)
  - Wattage
  - Style and features (e.g., twister, reflector, A-Lamp, globe, dimmable)
  - Manufacturer and product identifier (e.g., UPC or SKU codes)
  - Rated lumens
- Number of products incented (e.g., number of packs and bulbs per pack)
- Date incentive paid
- Dollar value of incentives paid
- Company name receiving incentive
- Location where products were ultimately 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)
- Assumptions regarding any parameters to savings estimates

[2] International Performance Measurement and Verification Protocol (IPMVP), which is considered the gold standard for evaluating energy-efficiency programs.

DRAFT

Similar details should be collected for programs using other delivery strategies. For example:

- Data collected for an audit program would include information about the date installed, the numbers and types of products installed, the wattage of the replaced bulb and location (room type), and contact information for the installation.
- Data collected for giveaway programs should contain at least the customer contact information and the quantity/type of product given away.

At a minimum, the evaluation should include a basic verification of savings, whereby the evaluator does the following: (1) sums up the detailed transactions, and (2) attempts to replicate the calculation of total claimed savings for the specific time period in which the savings were claimed, such as a program year or cycle.

Discrepancies between claimed and verified number of measures should be treated as adjustments to the number of program measures. In other words, if the *total* number of measures distributed does not match the number of measures claimed by a program administrator, the number of measures assumed sold or distributed should be adjusted accordingly. (That is, if the number of measures claimed by a program administrator does not match what is in the detailed tracking data, the tracking data should be regarded as correct.)

4.2 Delta Watts

Delta watts represent the difference between the wattage of the efficient lighting measure and the wattage of the assumed baseline measure. As noted, the wattage of the efficient measure should be available from the program tracking database. Where possible—such as with direct installation programs—the program implementation contractor should record the wattage of the particular lamp that the program measure is replacing.

Typically, this is done at the time of the audit, when the existing measure is replaced with the efficient measure. However, this is not possible for most program delivery strategies, so baseline wattage often needs to be estimated. In addition, the baseline assumptions need to incorporate the transition to EISA standards, beginning in 2012.

4.3 Approaches for Estimating Baseline Wattage

Recent studies have used a number of approaches for estimating baseline wattage, including these:

- ***Self-Report.*** This approach uses customer surveys after the installation to collect the wattage that was used before the energy-efficient lighting was installed.
- ***In-Home Inspections to Examine Wattage of Equivalent Fixtures.*** Using this approach, the implementation contractor examines the labeled wattage of bulbs in similar fixtures in each home to estimate the wattage that was used before the energy-efficient lighting was installed.
- ***Multipliers.*** This approach assumes that the baseline is a multiple—e.g., three to four times the wattage—of the efficient measure, so that one value (one multiplier) is used across all program bulbs.

- ***Manufacturer Rating.*** Most energy-efficient lighting products prominently list the replacement wattage assumptions on the box (see Figure 1). Manufacturers are also required to include detailed information regarding lamp output and efficacy as part of the "Lighting Facts" label that is now required on all retail lamp packaging. (http://www.ftc.gov/os/2010/06/100618lightbulbs.pdf)
- ***Lumen Equivalence.*** EISA standards include lumen ranges and assumptions regarding the equivalent wattage of incandescent lights.

Figure 1. Example of Manufacturer Rated Baseline Wattage

4.4 Recommended Approach

Each of these approaches has a number of strengths and limitations (see Table 1). Weighing each of these, the Residential Lighting Evaluation Protocol recommends using a lumen equivalency approach to estimate delta watts for conditions where the baseline wattage cannot be collected by the program implementation contractor at the time of measure installation. This approach is recommended because: (1) it provides consistency with the EISA requirements, and (2) most manufacturers' rated baseline wattage is already based on similar lumen categories.[3]

Alternatively, for studies that have sufficient budget to screen for a statistical sample of recent CFL purchasers, the self-report approach may be used to estimate delta watts (as well as other purchase attributes, including location and price). The UMP recommends, however, that the consumer recall approach apply these time limits:

[3] When the assumed baseline from the lumen equivalency approach differs from the manufacturer-rated baseline wattage, this is typically due to a lower-lumen bulb rated as a higher assumed baseline. For example, the manufacturer rates a bulb as a 120-W replacement, but the lumen output is more typical of a 75-W bulb. In these cases, consumers may "bin shift" up to a higher wattage of efficient product to get the light output they expect. Thus, the method recommends using the more conservative and lower assumed baseline wattage rather than what is printed on the box.

DRAFT

- A maximum of a six-month "window" (and preferably a three-month "window") for standard spiral CFLs, and
- Up to a year for specialty CFLs and LEDs, as these have far lower incidence but represent larger purchase decisions.

Note the self-report approach does offer the advantage of capturing consumer "bin-shifting."[4]

DRAFT

[4] A literature review did not reveal any studies that assess the magnitude of bin shifting, although forthcoming studies conducted by Navigant Consulting and the NMR Group found some evidence that customers purchased a higher-wattage bulb than the recommended replacement.

Table 1. Strengths and Limitations of Alternative Delta Watts Estimation Approaches

Approach for Estimating Baseline Wattage	Strengths	Limitations	Recent Studies Using Approach	Estimated Incandescent to CFL Ratio[5]
Customer self report	Capture customer intentions and bin shifting	Potentially low recall and social desirability bias	Duke Energy Residential Lighting Program (2010)	4.25
Examining equivalent fixtures	Actual recording of baseline wattage for existing measures	Difficult to truly identify equivalent fixtures; high cost to conduct statistically representative on-site study	2006-2008 California Upstream CFL Program	3.6
Standard multipliers	Low effort, 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.	Mid-Atlantic TRM (2011) Ohio TRM (2010)	3.95 4.25
Manufacturer rated baseline wattage	Widely available, relatively inexpensive to implement. Based off of wattage rating on package, often prominently displayed on the product	Some cases where the marketed baseline wattage exceeds the equivalent lumen output which may lead to "bin shifting"	WI Focus on Energy 2007 Residential Lighting Program	4.0
Lumen equivalence	Widely available, relatively inexpensive to implement. In most cases matches marketed baseline wattage, matches up with EISA standards	May provide conservative estimate in cases where marketed baseline wattage exceeds rated lumen output	ComEd PY3 Residential Lighting Program (forthcoming)	N/A

DRAFT

[5] The incandescent-to-CFL wattage will vary, based on both the types of bulbs promoted (e.g., standard vs. specialty) and the typical program CFL wattage. In addition, this ratio is sometimes shown as the ratio of the delta watts to CFL. (For example, the Mid-Atlantic TRM recommends a delta watts-to-incandescent ratio of 2.95).

Table 2 provides the assumed baseline wattage based on lumen range and incorporates the timing of EISA requirements as the new baseline standards.

Table 2. Estimated Baseline Wattage for Lumen Equivalencies[6]

Lumen Range	2011 Baseline	2012 Baseline	2013 Baseline	2014 Baseline
1490—2600	100-W	72-W	72-W	72-W
1050—1489	75-W	75-W	53-W	53-W
750—1049	60-W	60-W	60-W	43-W
310—749	40-W	40-W	40-W	29-W

While there may be "sell through" of existing product during the phase-in years, the Residential Lighting Evaluation Protocol recommends using the new baseline values for the entire year in which they take effect *unless* research shows significant "sell through" periods. (See the Uncertainty section later in this chapter).[7]

In addition, baseline wattage should be calculated for each lamp in the tracking database. The total estimated delta watts, therefore, is calibrated to the actual type and number of measures sold or distributed through the program.

There are two additional points of clarification for this approach:

- For lumens above or below these ranges, the marketed baseline wattage reported on the product should be used. In other words, lumens above the ranges in Table 2 might qualify for a 150-W baseline.
- EISA has a number of exceptions, including three-way bulbs, candelabras, and reflectors. In these cases, the baseline wattage should continue to be the 2011 standard incandescent wattage based on the lumen equivalence.

4.5 Replacements of Efficient Lighting Products With Newer Efficient Lighting Products

This methodology assumes that at the time of measure failure, the consumer has the choice of installing an energy-efficient lighting product or a standard-efficiency lighting product, regardless of what was previously installed. In areas with long history of CFL promotion—and as market penetration increases for CFLs or other high-efficiency lighting products —there is a higher probability that some fraction of the efficient-lighting products distributed through programs are being used to replace an installed CFLs that fail.

---

6 Shading represents initial year of EISA phase-in requirements.

7 EISA requires an even more efficient lighting standard in 2020 that is on par with current CFL efficacies. The life cycle savings of CFLs, therefore, should terminate for any remaining years beginning in 2020, and the life cycle savings for LEDs should incorporate this upcoming baseline change.

There are two approaches available to address this issue.

- The first is to assume the baseline is the federal standard (e.g., EISA), even if the consumer had previously had installed a CFL or LED. In this approach the CFL-to-CFL replacement scenario is assumed to be handled under investigation of program attribution, where it is more likely that consumers replacing CFLs with other CFLs may be freeriders.[8]
- The second is to revise the baseline wattage assumptions to reflect the share of in-kind replacement of CFLs. This approach requires the collection of data on the proportion of high-efficiency lamps distributed through the program that are replacing existing CFLs.

To avoid underestimating program savings, the UMP recommends that only one, rather than both, of these adjustments be applied. For jurisdictions that do not include any application of a net-to-gross adjustment, this would require using the second approach—conducting a market characterization study to determine the baseline and the percentage of high-efficiency lighting products that are replacing CFLs.

Finally, as more efficiency programs promote LEDs in the future, further research will be required to investigate the likelihood that energy-efficiency minded consumers are replacing CFLs with LEDs.

4.6 Uncertainty Regarding the Baseline and the Need for Ongoing Research

The recommended protocol acknowledges uncertainties around the residential lighting market in the next few years. These uncertainties deal with the types and prices of future lighting products that will be available on the market. Another source of uncertainties regards consumer reactions to the requirements and new products—for example, potential product hoarding, "bin jumping" to different incandescent wattage levels, and how quickly retailers sell through the existing product inventories.

The uncertainty around EISA was further heightened in December 2011 with the passage of the FY2012 Omnibus funding bill, which included a rider that halted funding for the Department of Energy to enforce the new standards. The National Electrical Manufacturers Association (NEMA), representing more than 95% of the U.S. lighting manufacturing industry, issued a press release after the passage of the bill stating that they did not support it. NEMA also points out

---

[8] The "New England Residential Lighting Markdown Impact Evaluation, January 20, 2009" found that 43% of respondents (24 out of 56) stated that the CFLs recently purchased and not installed were intended for use to replace incandescent lighting. That is, 57% of the respondents intended to use the stored CFLs to replace existing CFLs when they failed. While this was used to discount the delta watts, if those respondents who are already intending to replace CFLs with CFLs are presumably counted as freeriders, then program attribution should already incorporate any necessary adjustments.

that: (1) American manufacturers have invested millions of dollars in transitioning to energy-efficient lighting, and (2) EISA gave state attorneys general the authority to enforce the standards.

Thus, in cases where actual pre-program measure wattage is not available, the Residential Lighting Evaluation Protocol recommends that the EISA standards continue to be adopted as the new baseline. However, program administrators having adequate resources are recommended to conduct ongoing monitoring and research to determine whether the delta watts assumptions reflect actual market conditions during the phase-in of the EISA requirements. In particular, research in California—where the standards take effect one year in advance of the rest of the United States—may be informative for determining retailer and manufacturer reactions to EISA.

4.7 Annual Operating Hours

Hours-of-use (HOU) represents the estimated hours per year that the efficient lighting product will be used. Recent studies have shown a wide range of estimated HOU for CFLs, from a low of 1.5 to a high of 2.98 hours per day (see Table 3). A myriad of factors affect differences in the expected number of hours that efficient lighting products are used per year, including differences in demographics, housing types and vintages, CFL saturation, room type, electricity pricing, and even annual days of sunshine. As a result, extrapolation of data from one region has not proven successful in accounting for these influencing factors.[9]

Based on these disparate results, this protocol recommends that program administrators collect primary data through a metering study for residential lighting measures.

DRAFT

[9] For example, Cadmus' analysis of metered CFL hours-of-use, conducted as part of the evaluation of 2010 EmPOWER Maryland Residential Lighting and Appliances Program, revealed a significant difference in average daily hours of use as compared to extrapolating the hours-of-use from the ANCOVA model developed as part of the evaluation of the 2006-2008 California Upstream Lighting Program.

Table 3. Estimated CFL Hours of Use from Recent Metering Studies

Region	Publication Year	Author	Sample Size (Homes)	# of Efficient Bulbs Metered	Estimated Average Daily HOU
IL	Forthcoming	Navigant Consulting	67	527	2.7
Maryland (EmPOWER)	2011	The Cadmus Group, Inc./Navigant Consulting	61	222	3.0
CA (PG&E, SCE, and SDG&E service areas)	2006	KEMA, Inc.	375	983	2.3
CA (PG&E, SCE, and SDG&E service areas)	2010	KEMA, Inc. (The Cadmus Group, Inc., Prime Contractor)	≈1,200	N/A	1.9
CT, MA, RI, VT	2008	Nexus Market Research, Inc. et. al.	157	657	2.8
Pacific Northwest	2010	Northwest Regional Technical Forum, based on CA, 2010 KEMA, Inc.	N/A	N/A	1.9 for existing homes, 1.5 for new homes
Ohio	2010	Vermont Energy Investment Corporation (from Duke Energy)	N/A	N/A	2.8
North Carolina / South Carolina	2011	TecMarket Works and Building Metrics	34	156	2.5 (NC) / 2.7 (SC)

4.8 Metered Data Collection Method

Metering should be based on the following factors and associated guidelines, which are described in this section:

- Logger type
- Length of metering period
- Information collected on site
- Data integrity

4.8.1 Logger Type

Change-of-state loggers are preferred because they can capture short intervals and switch rates (the number of times lights are turned on and off). In addition, current-sensing meters (rather than light-sensing meters) are one approach for outdoor conditions in which ambient light can potentially inflate the estimated hours of use.

4.8.2 Length of Metering Period

Due to the seasonality of lighting usage, logging should: (1) be conducted in total for at least six months; and (2) capture summer, winter, and at least one shoulder season—fall or spring. At a

DRAFT

minimum, loggers should be left in each home for at least three months (that is, two waves of three-month metering will attain six months of data). All data should be annualized using techniques such as sinusoidal modeling to reflect a full year of usage.[10]

4.8.3 Information Collected On Site

In-home lighting audits should be conducted for all homes participating in the metering study. The audits should record the number and type of high-efficiency lighting products by fixture and room type. It is highly recommended that a full socket inventory be conducted to allow for an estimate of saturation of high-efficiency lighting equipment.

4.8.4 Data Integrity

All metered data need to be thoroughly cleaned to check for errant and erroneous observations.

4.9 Metering Sample Design

Ideally, metering is conducted for large samples of all major lighting types (including incandescent baseline lamps and fixtures); however, in practice, most evaluations do not have adequate resources for a scope of this size. Consequently, to optimize the allocation of moderate evaluation resources, target the metering to select lighting measures—typically CFLs—that represent the majority of savings in a residential lighting program. For measures representing a small percentage of savings (such as LEDs in more recent programs), the overall HOU should be estimated by examining the CFL hours-of-use for similar rooms and fixture types.

Given the difficulty of identifying program bulbs in an upstream program, loggers may be placed on 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 mark-down or buy-down. For homes that have many efficient lighting products, a subsample of fixtures may be selected, so long as they are selected randomly within the home. For example, if a home selected for a metering study has CFLs in 10 fixtures, meters can be placed on three to five randomly selected fixtures. This will both minimize the invasiveness in homes that are highly saturated with efficient lighting products and allow for a more cost-effective approach to include a larger sample of homes in the study.

The total number of loggers installed should be determined based on the desired levels of statistical confidence and precision, assuming a coefficient of variation based on recent studies of

---

[10] Sinusoidal modeling assumes that hours-of-use will vary inversely with hours of daylight over the course of a year. Sinusoid modeling shows that: (1) hours-of-use change by season, reflective of changes in the number of daylight hours and weather; and (2) these patterns will be consistent year to year, in the pattern of a sine wave. An example of this approach is provided in the evaluation of the 2006-2008 California Upstream Lighting Program evaluation.

programs with similar CFL saturation (using maturity of program as a proxy, if necessary) and housing characteristics.[11]

Following metering and annualization of results, the distribution of loggers by room type should be compared to the actual distribution of efficient lighting products per room type, as collected at the time of the audit. The hours-of-use should then be weighted 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 reveals that 15% of all CFLs are installed in kitchens, the data from the loggers in kitchens should be weighted up by 1.5 when estimating calculating total hours of use.

In addition, the demographic and household characteristics of the metering sample should be compared with the characteristics of the total population of homes believed to have purchased efficient lighting products. (This information can be collected through telephone surveys.) If significant differences appear *and* there is a large enough sample to support re-weighting based on such characteristics, the results should be weighted to reflect these differences.

4.10 Utilizing Secondary Data

While metering is the recommended approach, program administrators who are just launching a program—or do not have sufficient resources to conduct a metering study—may use secondary data from other metering studies.[12] This protocol recommends using the following criteria when selecting and using secondary data to use to estimate hours of use:

- Similarities in service territories
- Maturity of program or measure saturation
- Appropriate sample size
- Length of metering period
- Adjustments to reflect hours-of-use by room type

4.10.1 Similarities in Service Territories

Selecting a similar service territory based on geographic proximity and as many common demographic and household characteristics as possible will increase the likelihood that the secondary data provide a valid, reasonable, and accurate estimate.

---

[11] Recent Cadmus studies for Ameren Illinois and EmPower Maryland found CVs of approximately 0.6; however, the CV could be higher for mature programs where CFLs are in a wider selection of fixtures with more variable hours of use. Actual sample size should exceed the required number by approximately 5%-10% to allow for attrition due to data cleaning.

[12] As discussed under Section 7 of the Introduction chapter to the UMP Report, small utilities (as defined under the SBA regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies, such as that presented here, should be considered for such utilities.

4.10.2 Maturity of Program or Measure Saturation

Hours-of-use are expected to drop as the saturation of energy-efficient products increases, resulting in the installation of these products in less-used fixtures. Saturation is typically tied to the maturity of the program. In other words, regions with longer-running efficiency programs that have higher saturation rate are expected to have lower hours-of-use.[13] Using secondary data from programs of similar maturity levels will increase the data's applicability.

4.10.3 Sample Size

The number of observations varies considerably between studies, so the sample size, standard errors, and precision levels at equivalent confidence levels should be compared across studies.

4.10.4 Length of Metering Period

Studies that capture both winter and summer usage may be more appropriate for estimating overall annual use.

4.10.5 Adjustments to Reflect Hours-of-Use by Room Type

Extrapolating data from one region to another should be conducted by calibrating to the different levels of measure saturation by room type. If possible, the hours-of-use by room type from a secondary data source should be weighted by the room type distribution of CFLs for the region under study.

4.11 Snapback/Rebound or Conservation Effect

"Snapback" or "rebound" refers to changes in use patterns that occur after the installation of an energy-efficient product and result in reducing the overall measure savings. For example, when residential lighting customers use a CFL for more hours per day than they used the replaced incandescent bulb, this constitutes snapback. This behavior change may be due to factors such as the cost savings per unit of time from the CFL or a concern that turning CFLs on and off shortens their effective useful life (although it is unlikely most consumers are aware of this effect on life). Some customers, however, might have lower hours of use after installing a CFL, perhaps due to a corresponding desire to reduce energy consumption.

Due to the nature of residential lighting programs, it is not typically possible to conduct metering both before and after installation of energy-efficient lighting. Therefore, the Residential Lighting Protocol does not recommend adjusting for snapback/rebound effects in the hours-of-use estimates.[14]

---

[13] For example, hours-of-use in California dropped from an average of 2.3 hours per day in the 2004-2005 program year study to 1.9 hours per day in the 2006-2008 program year study. CFL socket penetration (the percentage of sockets containing CFLs) increased from 9% in the 2004-2005 study to 21% in the 2006-2008 study.

[14] While surveys can be used to estimate potential snapback behavior, these efforts are considered more qualitative. Also, surveys cannot easily capture the relationship of hours-of-use between multiple fixtures. For example, after a retrofit, a home owner may consciously choose to use a fixture for more hours—rather than a standard-efficiency fixture—as a strategy to save additional energy).

DRAFT

4.12 In-Service Rate

The in-service rate represents the percentage of incented residential lighting products that are ultimately installed by program participants. In-service rates 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 *and* the customer may not have requested the more-efficient lamps.

For upstream programs, three factors—as shown in Table 4, have led to first-year, in-service rates well below 100%: (1) the often deeply discounted price, (2) the inclusion of program multipacks, and (3) the common practice of waiting until a bulb burns out before replacing it.

DRAFT

Table 4. Estimated First-Year, In-Service Rates from Recent Evaluations of CFL Upstream Lighting Programs

Region	Publication Year	Author	Percentage of CFLs Installed in Program Year*
Arizona (APS service area)	2008	Navigant Consulting	90%
California (PG&E, SCE, and SDG&E service areas)	2010	KEMA, Inc. (The Cadmus Group, Inc., Prime Contractor)	67%
Conn., Mass., R.I., Vt.	2009	Nexus Market Research Inc., et al.	76%
Illinois	2012 (Forthcoming)	Navigant Consulting and Itron, Inc.	71%

*Based on program year only, not years subsequent to the program year or several years in a multiyear program cycle.

The Residential Lighting Protocol recommends that in-service rates be estimated using different methods, as determined by the delivery mechanism:

- For ***direct installation programs***, conduct verification (such as, telephone survey or site visits) to assess installation and measure persistence, regardless of whether working bulbs were removed before they failed.
- For ***giveaway or coupon programs***, conduct verification when customer contact information is available. Also, ask respondents whether: (1) the installation location was within the relevant service territory, and (2) the measure was installed in a home or 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 for a similar program where customer information was collected) or, if necessary, on the in-home audit approach (described under the next bullet).
- For ***upstream programs***, calculate in-service rates through an in-home audit. Since program bulbs cannot be easily identified, the in-service rate can be calculated 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 the recent 12-month period is insufficient to provide the necessary levels of confidence and precision, apply a long-term, in-service rate can be using all bulbs, regardless of the time of purchase.
- Although the in-home audit is the recommended approach, a telephone survey can be utilized when program administrators are just ***launching a program or do not have sufficient resources*** to conduct an in-home audit. To minimize recall bias, the callers should focus questions only on products purchased in the recent 12-month period rather than the period covering the long-term, in-service rate. (Studies have shown that respondents tend to have better recall about the percentage of bulbs purchased and installed within the past 12 months, as compared to the percentage of bulbs that have ever been purchased and installed.)

DRAFT

Although first-year, in-service rates for upstream programs are less than 100%, recent studies have demonstrated that consumers plan to install virtually all of the incented bulbs; however, they sometimes wait until an existing bulb burns out.[15] As a result, program administrators have been able to take credit in one of two ways for savings that occur in years following the year that the incentive was paid:[16]

- ***Discount Future Savings.*** In this method, all of the costs and benefits are claimed during the program year, but the savings from the expected future installation of stored program bulbs are discounted back to the program year using a societal or utility discount rate.
- ***Stagger Timing of Savings Claims.*** In this method, all of the expenses are claimed during the program year, but the savings (and, therefore, the accompanying avoided cost benefits) are claimed in the years in which the program measures are estimated to be installed.

To calculate the installation rate trajectories, the Residential Lighting Evaluation Protocol recommends using the findings from the evaluation of the 2006-2008 California Residential Upstream Lighting Programs, which estimated that 99% of program bulbs get installed within three years, including the program year.[17] Because the study examined three years of program activity, it does not specifically include the percentage of bulbs installed by the year following program activity; it only estimates the total after three years. Therefore, program administrators should assume the bulbs that will be installed in future years are split equally between one and two years following the program year, calculated as:

$$ISR_{PY2} = \frac{99\% - ISR_{PY1}}{2}$$

$$ISR_{PY3} = \frac{99\% - ISR_{PY1}}{2}$$

As noted in the delta watts discussion, this methodology does not adjust for CFL-to-CFL replacement, which is assumed to be handled by assessments of program attribution.

---

15 For example, the evaluation of the Program Year 2 Commonwealth Edison Residential Energy Star Lighting Program found that about of customers 90% were waiting until a working incandescent or CFL burned out before they installed the stored CFLs (Table 3-6).

16 The selection of which approach to use will depend upon the study purpose and regulatory requirements.

17 Few studies have attempted to quantify installation rate trajectories, and these protocols recommend this as an additional area for further research.

4.13 Interactive Effects with Heating, Ventilation, and Cooling

CFLs and LED lamps give off less waste heat than do incandescent bulbs, which affects heating, ventilation, and cooling (HVAC) energy requirements. These effects vary based on space conditioning mode (and saturation) and climate zones. The interactive effects depend on a variety of house-specific factors, and the net impact on lighting energy cost savings could be positive, negative, or neutral.[18] In cooling-dominated climates, the interactive effects are positive, resulting in additional savings due to decreased cooling load. However, in heating-dominated climates, the interactive effects are negative, with decreased savings due to increased heating load.

Because of the potential impacts of interactive effects, the Residential Lighting Evaluation Protocol recommends these effects be included in evaluations of residential lighting programs.[19] One approach is to estimate these effects through the use of 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, secondary sources—such as the Residential Energy Consumption Study (RECS) —can be used to estimate these inputs. Other recent approaches include a billing analysis.[20]

Some regions have developed tools based on such simulations (for example, in California, the Database for Energy Efficiency Resources [DEER][21] and the Regional Technical Form [RTF] in the Northwest. Such regional collaboration can minimize the cost of determining the interactive effects for those 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, this recommends using a value from an existing resource, but ensuring that at least the climate (heating and cooling degree days) and, ideally, the latitude, HVAC system types, and saturations are similar between the program administrator's territory and the territory from which the data are taken.

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

---

[18] Parkeh, Anil. "Do CFLs Save Whole-House Energy?" Home Energy Magazine, November/December 2008, pages 20-22. See also Parekh, A., M. C. Swinton, F. Szadkowski, M. Manning, 2005, "Benchmarking of Energy Savings Associated with Energy Efficient Lighting in Houses", National Research Council Canada. NRCC-50874. (Can also be downloaded at: http://nparc.cisti-icist.nrc-cnrc.gc...n&req=20377557)

[19] Note that interactive effects are only relevant for bulbs installed in conditioned spaces. Thus, exterior lights will not have HVAC interactive effects.)

[20] Brunner, Eric J., Peer S. Ford, Mark A. McNulty, Mark A. Thayer, Compact Fluorescent Lighting and Residential Natural Gas Consumption: Testing for Interactive Effects, Energy Policy 38 (2010), 1288-1296.

[21] www.deeresources.com/DEER2011/download/LightingHVACInteractiveEffects_13Dec2011.xls

administrator's service territory. Therefore, some program bulbs may go to non-residential customers or to customers served by other utilities. These parameters are discussed in this section.

5.1 Cross-Customer Class Sales

Non-residential customers typically use lighting products for more hours per day than do residential customers. Typically, non-residential customers also have higher peak coincidence factors. Therefore, sales of efficient lighting products to non-residential customers may lead to higher savings than those assumed through the methods outlined above.

The typical approach to estimating this parameter has been through customer intercept surveys, where customers who purchase lighting products participate in a short survey—asking about intended installation location and facility type—at the time of sale. This parameter has also been estimated through surveys with store managers (asking them to estimate the percentage of bulbs sold to non-residential customers) or with the owners of small businesses (asking them where they typically purchase lighting products).

The Residential Lighting Evaluation Protocol recognizes several key limitations in estimating this parameter, including:

- ***Customer intercepts may not represent all program sales***. Conducting customer intercept surveys can be expensive, and they are typically conducted only in high-volume stores (such as Home Depot, Lowes, and Wal-Mart, etc.). In some cases, these surveys are conducted only during high-volume promotions. Also, since some retailers refuse to allow the surveys to be conducted, the surveys may not be representative of total program sales.

  Accuracy from intercepts is further challenged because business owners and contractors: (1) may be a minority of purchasers, (2) may purchase more units per visit than residential purchasers, and (3) may not purchase during the same time as the average residential purchaser.

- ***Surveys lack high reliability***. Store managers usually do not have detailed information on program bulb purchasers, so their estimates of sales to non-residential customers may be unreliable. Surveys of small business customers also face challenges, as there is nonresponse bias (that is, calling a small business and not getting cooperation from the business decision maker to take a survey). Additionally, quantifying the number and type of bulbs purchased by channel may have recall bias.

5.2 Cross-Service Area Sales (Leakage)

Recent studies have also attempted to estimate the number of program bulbs sold to customers outside of the program administrator's service territory. This is commonly referred to as "leakage" or "spillage."

This protocol recognizes several key limitations in estimating this parameter, including:

- ***Cross-Region Sales.*** Many neighboring service territories are now targeted by residential lighting programs; thus, there is less of an incentive to shop outside one's own service territory to purchase less-expensive lighting products. In some cases, leakage of program bulbs occurs in both directions across service territory boundaries, which may offset the effect in either or both territories.
- ***Many programs now limit participating retailers, so that 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 population of program customers as to minimize potential leakage.

5.3 Estimating Cross-Customer Class and Cross-Service Area Sales

Based on the limitations of estimating these parameters—and the fact these parameters may offset each other (that is, the increased savings of sales to non-residential customers may be at least partially offset by leakage) —this protocol recommends excluding these parameter estimates from impact evaluations of upstream residential lighting programs.

For program administrators who are using intercepts for other purposes (including an assessment of program attribution), questions regarding the intended location and business type can be included in surveys. However, the results should be used cautiously with the following adjustments:

- The results should be weighted to reflect the percentage of program bulbs represented by those distribution channels. For example, if intercept surveys are conducted at retailers that represent 75% of program bulbs, the findings should be assumed to reflect 75% of program bulb sales. For those distribution channels that have not received intercept surveys, the evaluator should first assess how the cross-customer class and cross-service area sales might differ and then apply extrapolated values.
- Intercept surveys should be conducted at retailer storefronts that represent a mix of likely leakage (based on the distance to adjacent service territories). Alternatively, the results should be weighted to reflect the actual mix of retailer risk of leakage.

6 Program Evaluation Elements

Residential lighting programs offer a variety of measures through multiple delivery strategies, with the upstream CFL programs currently being the most ubiquitous. Program administrators who offer a variety of measures and rely on a variety of delivery strategies may need to prioritize their evaluation resources based on criteria such as contribution to savings and assessed uncertainty.

Savings should be assessed through a mix of primary and secondary data, using IPMVP Option A (Retrofit Isolation: Key Parameter Estimates). Key areas needing ongoing and additional research are these:

- Assumptions regarding baseline wattage as EISA standards take effect and as LEDs become a larger source of program savings. (For example, customers who would have installed a CFL, rather than a program-incented LED, in absence of the program).
- Installation trajectories for measures that are not installed in the first year.

DRAFT

7 Bibliography

ASHRAE Performance Measurement Protocols for Commercial Buildings, ASHRAE 2010.

ASHRAE, Research Project 1093 Compilation of Diversity Factors and Schedules for Energy and Cooling Load Calculations, ASHRAE Research Report, 2000.

ASHRAE Guideline 14-2002 Measurement of Energy and Demand Savings, ASHRAE, 2002.

ASHRAE Guideline 14-2002R (Revision of Guideline 14, currently in process, publication date TBD).

Brunner, Eric J., Peer S. Ford, Mark A. McNulty, Mark A. Thayer, Compact Fluorescent Lighting and Residential Natural Gas Consumption: Testing for Interactive Effects, Energy Policy 38 (2010), 1288-1296.

Cadmus Group, Inc., Lighting and Appliance Evaluation – PY 2, Prepared for Ameren Illinois, December 2010.

Glacier Consulting, Analysis of Delta Watts Values for CFLs Rewarded through the Residential Lighting Program during FY07, prepared for the State of Wisconsin Public Service Commission, March 6, 2008.

KEMA, Inc. and Cadmus Group, Inc., 2010. Final Evaluation Report: Upstream Lighting Program. Prepared for California Public Utilities Commission, Energy Division. February 28, 2010.

KEMA, Inc., CFL Metering Study. Prepared for California's Investor-Owned Utilities (PG&E, SCE, SDG&E, and SoCalGas), February 2005.

Mid-Atlantic Technical Reference Manual, Version 2.0, Prepared by Vermont Energy Investment Corporation, Facilitated and Managed by Northeast Energy Efficiency Partnerships, July 2011.

Navigant Consulting and Itron, Plan Year 2 Evaluation Report: Residential Energy Star Lighting. Prepared for Commonwealth Edison Company, December 21, 2010.

Navigant Consulting and Cadmus Group, Inc., EmPOWER Maryland 2010 Interim Evaluation Report. Prepared for Baltimore Gas and Electric, Potomac Electric Power Company, DPL, SMECO, and AP, January 15, 2011.

Nexus Market Research, Inc. and RLW Analytics, Inc. Impact Evaluation of the Massachusetts, Rhode Island, and Vermont 2003 Residential Lighting Programs. Prepared for The Cape Light Compact, State of Vermont Public Service Department for Efficiency Vermont, National Grid, Northeast Utilities, NSTAR Electric, Unitil Energy Systems, Inc. October 2004.

DRAFT

Nexus Market Research, Inc., RLW Analytics, Inc., and GDS Associates, 2009. Residential Lighting Markdown Impact Evaluation. Prepared for Markdown and Buydown Program Sponsors in Connecticut, Massachusetts, Rhode Island, and Vermont. January 20, 2009.

PA Consulting Group and NMR Group, Focus on Energy Evaluation 2009 CFL Savings Analysis, April 22, 2010.

Parekh, A., M. C. Swinton, F. Szadkowski, M. Manning, 2005, "Benchmarking of Energy Savings Associated with Energy Efficient Lighting in Houses", National Research Council Canada. NRCC-50874.

Parkeh, Anil. "Do CFLs Save Whole-House Energy?" Home Energy Magazine, November/December 2008, pages 20-22.

State of Ohio Energy Efficiency Technical Reference Manual, Prepared for the Public Utilities Commission of Ohio by Vermont Energy Investment Corporation, August 6, 2010.

TecMarket Works and Building Metrics, Duke Energy Residential Smart $aver CFL Program in North Carolina and South Carolina: Results of a Process and Impact Evaluation, February 15, 2011.

DRAFT

Refrigerator Recycling Evaluation Protocol

Doug Bruchs, The Cadmus Group, Inc.

Refrigerator recycling programs have become a staple of residential demand-side management portfolios.

1 Measure Description

Refrigerator recycling programs are designed to save energy through the removal of old-but-operable refrigerators from service. By offering free pick-up, providing incentives, and disseminating information about the operating cost of old refrigerators, these programs are designed to encourage consumers to:

- Discontinue the use of secondary[1] refrigerators;
- Relinquish refrigerators previously used as primary units when they are replaced (rather than keeping the old refrigerator as a secondary unit); and
- Prevent the continued use of old 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 the recycling of the plastic, metal, and wiring components.

2 Application Conditions of Protocol

These brief descriptions indicate the range of designs currently seen in recycling programs:

- Some recycle both primary and secondary refrigerators.
- Some accept only secondary refrigerators.
- Some impose restrictions on vintage eligibility.
- Some are offered in conjunction with point-of-sale rebates to encourage the purchase of ENERGY STAR®-rated refrigerators.
- Some are 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

---

[1] Secondary refrigerators are units not located in the kitchen.

[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 the Savings Calculations section.

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]

3 Savings Calculations

The total gross energy savings[4] (kWh/year) achieved from recycling old-but-operable refrigerators is calculated using the following general algorithm:

Equation 1

$$GROSS_kWh_{SAVED} = N * UEC_{EXISTING} * PART_USE$$

Where:

$GROSS_kWh_{saved}$	=	Annual electricity savings measured in kilowatt hours
N	=	The number of refrigerators recycled through the program
$UEC_{EXISTING}$	=	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

The total net energy savings (kWh/year) is then calculated as follows:

Equation 2

$$NET_kWh_{SAVED} = GROSS\ kWh_{saved} * (1 - FR_RATIO) - (N * UEC_{REPLACEMENT} * PART_USE_{REPLACEMENT} * INDUCED_REPLACEMENT_RATIO)$$

Where:

FR_RATIO	=	Freeridership ratio (the proportion of gross savings that would have occurred in the program's absence)
$UEC_{REPLACEMENT}$	=	The annual unit energy consumption of the average replacement unit

[3] As discussed under Section 7 of the Introduction chapter to the UMP Report, small utilities (as defined under the SBA regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

[4] The evaluation protocol methods focus on energy savings; they do not include other parameter assessments such as peak coincidence factor (demand savings), incremental cost, or measure life.

DRAFT

$PART_USE_{REPLACEMENT}$ = The portion of the year the average replacement refrigerator is likely to operate

INDUCED_REPLACEMENT_RATIO = The proportion of participants reporting that they purchased a replacement refrigerator as a result of participating in the program[5]

The recommended techniques for estimating each of these parameters are described below.

4 Gross Savings

This section provides instructions for determining the parameters required to estimate a refrigerator recycling program's total gross savings ($GROSS_kWh_{SAVED}$).

The key parameters are these:

- Measure Verification (N)
- Annual Energy Consumption ($UEC_{EXISTING}$)
- Part Use Factor (PART_USE)

4.1 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, 90% level of confidence with a 10% margin of error. Past evaluations have shown that participants typically

[5] That is, the program caused customers to buy a new unit when they otherwise would not have. More information regarding induced replacement is included in this protocol's Net Savings section.

have little difficulty confirming the number of units recycled and the approximate date the removal took place (Cadmus 2010).

4.2 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 collected as part of other recycling program evaluations that occurred within the previous five years (when evaluation resources do not support primary data collection).

Deemed 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) utilizing the results to estimate annual energy consumption and, subsequently, energy savings.

4.2.1 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).[6] However, recent evaluations indicate that DOE test conditions (e.g., 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 utilized *in situ* (meaning "in its original place") metering to assess energy consumption.

In situ metering is recommended for two reasons:

- It factors in environmental conditions and usage patterns within participating homes (e.g., 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 publically 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 making a similar type of adjustment or assumption is unnecessary.

---

[6] Evaluations have also used forms of billing analysis; however, the 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 could be obtained from end-use specific approach.

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 old-but-operable appliances.

4.2.1.1 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 the Sample Design and Data Development chapter. 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.[7] For refrigerators, these attrition issues may include simple meter failure, relocation of the unit during metering, and atypical usage (e.g., the refrigerator is prematurely emptied in preparation for program pick-up). 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 the majority of recycled appliances would have been used as secondary units had they not been decommissioned through the program.[8] However, some units may continue to operate as a primary unit within the same home. To account correctly for differences in usage patterns between the usage type categories (e.g., primary and secondary refrigerators), it is critical to stratify the metering sample to represent the different usage types.[9]

  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, since 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 level of collinearity between refrigerator age, size, and configuration. However, should sufficient evaluation resources be available, targeting a sample of appliances with less-common

---

[7] For a broader discussion of the coefficient of variation see the, Sample Design and Data Development chapter.

[8] This includes several scenarios: The refrigerator may continue as a secondary appliance within the same home, be transitioned from a primary to secondary appliance within the same home, or become a secondary unit in another home.

[9] 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).

characteristic profiles 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.[10] 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), but the duration needs to be balanced with the customers' desire to have the 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, shorter intervals (of one or two minutes) are 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, and
    - Frequency and duration of door openings.[11]

  Not all of 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.[12] 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

DRAFT

---

[10] The previously cited evaluations in California (ADM, April 2008 and Cadmus, February 2010) both collected metering data for a minimum of from 10 to 14 days.

[11] The previously cited evaluation (Cadmus, February 2010) employed 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).

[12] Forthcoming *Michigan Energy Efficiency Measure Database* memo by Cadmus regarding Consumers Energy and DTE Energy appliance recycling programs.

(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 pick-ups), 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.

  Once participants are recruited, the evaluator and the implementer should collaborate in scheduling the participants' pick-up after all of 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 first wave the proper installation of metering equipment at a small sample of participating homes Thus, any installation issues can be identified and corrected.

  Since 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.*** Since 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.

4.2.2 About Regression Modeling

To estimate the annual UEC of the average recycled refrigerator, this protocol recommends that evaluators use a multivariate regression model(s) that relates observed energy consumption to refrigerator characteristics.

Evaluators should employ models that utilize 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's tracking database.

In estimating UEC, both time and cross-sectional effects must be accounted for. This can be done one of two ways:

- ***Use model that estimates simultaneously the impacts of time and cross-sectional effects on energy consumption.*** This approach is recommended if there are both a reasonably large sample size *and* observations of units across both summer and winter peak periods. Or,

- ***Use a set of time-series models.*** If metering occurred during only one or two seasons, use an existing refrigerator load shape to extrapolate the annual UEC for each metered refrigerators. Then use a second regression model utilizing 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 be used. 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/secondary designation
- Conditioned/unconditioned space[13]
- Location (kitchen, garage, basement, porch, etc.)
- Weather (cooling degree days [CDD] and/or heating degree days [HDD])

For each set of potential independent variables, evaluators should assess the variance inflation factors, adjusted R^2s, residual plots, and other measures of statistical significance and fit.

[13] Primary/secondary and conditioned/unconditioned space variables may exhibit a strong collinearity. Consequently, do not include both in the final model.

In the specification process, evaluators should also consider the following elements:

- Estimating model parameters by using an Ordinary/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 space 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.

The following sample regression model is based on data from 472 refrigerators metered and recycled through five utilities:

$$\begin{aligned} Existing\ UEC \\ &= 365.25 * [0.582 + 0.027 * (22.69\ years) + 1.055 \\ &* (63\%\ manufactured\ before\ 1990) + 0.067 * (18.92\ ft.^3) - 1.977 \\ &* (6\%\ single\ door\ units) + 1.071 * (25\%\ side - by - side) + 0.605 \\ &* (36\%\ primary\ usage) + 0.02 * (2.49\ unconditioned\ CDDs) - 0.045 \\ &* (1.47\ unconditioned\ HDDs)] \end{aligned}$$

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. Table 1 provides an example of this process, using average values for each independent variable from an example program.

DRAFT

Table 1. Example UEC Calculation Using Regression Model and Program Values

Independent Variable	Estimate 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 the program)	0.6054	0.36
Interaction: Located in Unconditioned Space x CDDs	0.020	2.49
Interaction: Located in Unconditioned Space x HDDs	-0.045	1.47
Estimated UEC (kWh/Year)		**1,240**

4.2.3 Utilizing 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 one that is 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 when: (1) *in situ* metering is not an option, *and* (2) a recently developed model from a single comparable program cannot be identified.

4.3 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, because:

- The UEC model yields an estimate of annual consumption, and
- Not all recycled refrigerators would have operated year-round had they not been decommissioned through the program.

Table 2 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 2. 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 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-time* were likely to have been operated. The protocol recommends this assessment be handled through the following multi-step process.

1. ***Ask participants where the refrigerator was located for the majority 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 type 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 all other locations as secondary.

   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 pick-up.

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 all primary units are operated year-round.)

3. ***Ask those participants who that their secondary refrigerator was operated for only a portion of the preceding year to estimate the total number of months during that time 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.

These three steps enable evaluators to obtain important and specific information as to how a refrigerator was used before it was recycled. The example program provided in **Error! Reference source not found.** shows that:

- The participant survey determined that 93% of recycled refrigerators were operated year-round either as primary or secondary units. (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.

DRAFT

- The remainder (3%) was operational for a portion of the year. Specifically, the survey determined that part-time refrigerators were operated for an average of three months a 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 utilizes the UEC determined through the methods described in Section 4.2.2. In this example, the program's secondary-only part-use factor is 0.88, while the overall part-use factor is 0.93.

Table 3. Example Calculation of Historical Part-Use Factors

Usage Type and Part-Use Category	Percent 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%**	**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%**	**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 4 demonstrates how a program's part-use factor is determined using a weighted average of historically observed part-use factors and participant's likely action in the absence of the program.[15] Here, the result is a part-use value of 0.90, based on the expected future use of the refrigerators had they not been recycled. Applying this value to the determined UEC (1,240 kWh/year) yields the program's average per-unit gross savings—in this case, 1,161 kWh/year.

[14] While the future usage type of discarded refrigerators is unknown, given the typical age and condition of participating appliances, it is most likely the refrigerator would have been used as a secondary unit after being transferred. While some transferred units may become primary appliances, applying the part-use factor for secondary-only units ensures part-use is not overestimated.

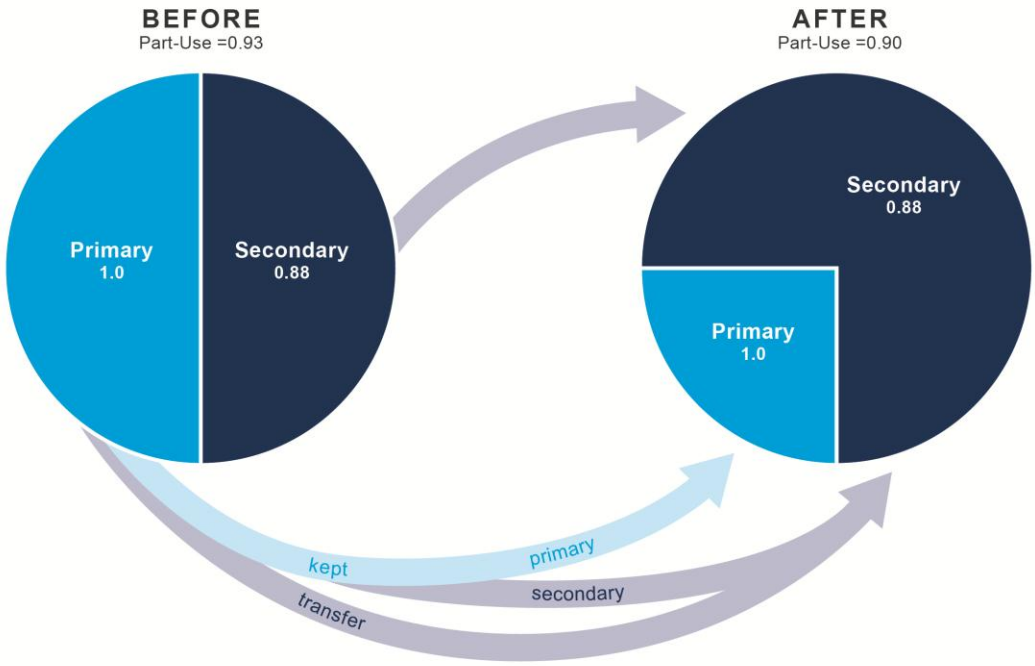
[15] Evaluators should not calculate part-use using participant's estimates of future use had the program not been available. Historical estimates based on actual usage rates are more accurate, especially since it is possible participants will underestimate future usage (believing they will only operate it part of the year, despite the fact the majority of refrigerators operate continuously once plugged in).

Table 4. Example Calculation of Prospective Program Part-Use

Use Prior to Recycling	Likely Use Independent of Recycling	Part-Use Factor	Percent of Participants
Primary	Kept (as primary unit)	1.0	15%
	Kept (as secondary unit)	0.88	25%
	Discarded	0.88	15%
Secondary	Kept	0.88	30%
	Discarded	0.88	15%
Overall	**All**	**0.90**	**100%**

The determination and application of part-use described above is summarized graphically in Figure 1.

Figure 1. Part-Use



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 part-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, since it may change more rapidly during the early stages of a program's life cycle. After a program has been in operation for at least three years, it is sufficient to conduct a part-use assessment every other year.

4.4 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 old 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, since 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[16] 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 operate once they are replaced.

However, when a recycling program induces replacement (i.e., the participant would ***not*** have purchased the new refrigerator in absence of the recycling program), evaluators must account for replacement. This issue is addressed in the following Net Savings section, which also discusses recycling program's impact on the secondary market and how evaluators should account for these effects. This protocol focuses on the actions of would-be recipients of refrigerators recycled through the program (that otherwise would have been transferred to a new user) when the recycled unit is not available.

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 evaluation's 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 also discussed in the following Net Savings section in more detail.

[16] With the exception of induced replacement, which is addressed in the Net Savings section.

5 Net Savings

This section provides instructions for determining the additional parameters required to estimate a refrigerator recycling program's net savings (NET_kWh_{saved}). 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).

The key additional parameters detailed in this section are these:

- Freeridership (FR_RATIO)
- Annual Energy Consumption of Replacement Refrigerator ($UEC_{REPLACEMENT}$)
- Induced Replacement Ratio (INDUCED_REPLACMENT_RATIO)

5.1 Spillover

For the following reasons, this protocol does not recommend quantifying and applying participant spillover to adjust net savings:

1. Unlike a CFL program, the opportunities for "like" spillover (the most common and defensible form of evaluated spillover for most downstream DSM programs) are limited with a recycling program as there is a limited number of refrigerators active in a home and available for recycling.

2. 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 "unlike" spillover.

3. The accurate quantification of spillover is challenging and, despite well-designed surveys, uncertainty often exists regarding the attribution of subsequent efficiency improvements to participation in the recycling program.

As a result of the ease of participation and high levels of participant satisfaction, appliance recycling programs may, however, 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.

5.2 Data Sources

After determining a program's gross energy savings, the net savings are determined by applying a NTG adjustment using the follow data sources[17]:

- ***Participant Surveys***. Surveys with a random sample of participants offer self-report estimates regarding whether participating refrigerators would have been kept or

[17] When it is cost-prohibitive to survey nonparticipants and interview market actors, calculate freeridership using participant surveys and secondary data from a comparable set of market actors.

discarded independent of the program.[18] When participants indicate 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 used-appliance dealer.

- ***Nonparticipant***[19] ***Surveys***. To mitigate potential response bias[20], 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.[21] Since nonparticipant surveys require greater evaluation resources, it is acceptable to use smaller sample sizes.[22]
- ***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 that requiring follow-up include:
    - "I would have sold it to a used appliance dealer," or
    - "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 so, only some are resold, what are characteristics (e.g., 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 participants' self-reported hypothetical actions independent of the program. This information can also be used to prompt participants to offer alternative hypothetical actions.[23]

---

18 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.

19 "Nonparticipants" are defined as utility customers who disposed of an operable refrigerator outside of the utility program while the program was being offered.

20 See the Sampling chapter for a broader discussion of sources of bias.

21 Information regarding the likelihood that the recycled refrigerator would have been retained independent of program intervention can be obtained reliably through the participant surveys.

22 The cost of identifying nonparticipants can be minimized by adding the nonparticipant NTG module to concurrent participant surveys for other utility program evaluations.

23 More detail is provided in Section 5.3 Freeridership (FR_RATIO).

A detailed explanation of how to estimate NTG by aggregating information from these sources is provided later in this section. Also, as previous recycling evaluations have found little evidence of program-induced spillover[24], this protocol does not require that spillover be addressed quantitatively.[25] As a result, estimates of NTG need only to account for freeridership and induced replacement.

5.3 Freeridership (FR_RATIO)

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 destruction.

5.3.1 Participant Self-Reported Actions

To determine the percentage of participants in each of the three scenarios (and, therefore, to calculate the program NTG), 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's market research revealed used appliance dealers were unlikely to purchase it (due to its age or

---

[24] See the Net-to-Gross chapter for more information about spillover.

[25] This issue is discussed further in Cadmus' forthcoming evaluation of PacifiCorp's Appliance Recycling Program in Washington.

condition), then the participant should be asked what they would have likely done *had they been unable to sell the unit to a dealer*. Evaluators should then use their response to this question to assess freeridership.

If market research determined local waste transfer stations charge a fee for dropping off refrigerators, then participants who initial specify this as their option should be informed of that fee. Then, ask the participants to confirm what they would have done in the absence of the program. Again, evaluators should this response to assess freeridership.

Use this iterative approach with great care. It is critical that evaluators find the appropriate balance between increasing the plausibility of participants' stated action (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 freeridership.

- Some final responses clearly indicate freeridership, such as: "I would have taken it to the landfill or recycling center myself."
- Other responses clearly indicate no freeridership, 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").

5.3.2 Impact on Secondary Market

If it is determined that the participant would have transferred the unit to another customer on the grid, the question becomes what that potential acquirer did. There are three possibilities:

1. ***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 prevailing program theory and dictates that participating appliances are essentially luxury goods for would-be acquirers. (That is, a secondary refrigerator is not a necessity, but it is nice to have should it be available.)
2. ***All of the would-be acquirers would find another unit.*** Thus, program participation has no effect on the total number of refrigerators operating on the grid. In this case, none of the energy consumption associated with avoided transfers should be credited to the program, as the total refrigerator load operating on the grid is essentially unchanged. 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.
3. ***Some of would-be acquirers would find another unit, while others would not.*** In this case, some of these 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).

Absent a detailed accounting of all metrics in the appliance market (rarely available), this is extremely difficult to measure. The identification of would-be acquirers is also problematic, as it is uncertain whether the units they acquired are of the same vintage and condition as participating appliances.

In some cases, evaluators have conducted in-depth market research to ascertain the appropriate attribution of savings for this scenario. Although these studies are imperfect, they provide a more accurate accounting of freeridership and, where feasible, this is the recommended approach. Unfortunately, this type of research is often cost-prohibitive or the necessary data are simply unavailable.

As a result, this protocol recommends the third possibility listed above—to assume that half (0.5, the midpoint of possibilities 1 and 2) of the would-be acquirers of avoided transfers found another unit instead.

5.3.3 Calculating Freeridership

To estimate freeridership, this protocol recommends that evaluators use the stated intentions of surveyed participants vetted against market intelligence collected through retailer interviews as described above. The protocol also recommends estimating a second, nonparticipant-based freeridership ratio, using information obtained from nonparticipants regarding how they actually discarded their refrigerator independent of the utility program. Nonparticipant responses should be categorized in a similar manner to participant responses and then be identified as either indicative or not indicative of freeridership, based on whether the appliance remained operational or was destroyed.

Evaluators should then calculate the program's overall NTG ratio as a weighted average of the freeridership ratios of both participants and nonparticipants. Participant and nonparticipant values are averaged to mitigate potential biases in each group.[26] As the true population of nonparticipants is unknown, the NTG values should be weighted using the inverse of the variance of participant and nonparticipant freeridership ratios.[27] This method of weighting gives greater weight to those values that are more precise or less variable. This approach is detailed in Table 5.

---

[26] 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 absence of the program due to participants self-selecting into the program (as opposed to being randomly enrolled)

[27] Inverse variance weights involve weighting each estimate by the inverse of its squared standard error ($1/SE^2$). This technique is common in the meta-analysis literature and is used to place greater weight on more reliable estimates.

Table 5. NTG Calculation Details

Survey Sample	Scenario	Proportion of Survey Sample	Discard Type	Proportion of Scenario	Scenario FR	FR*Proportion of SS*Proportion of Scenario	Sample FR	SE(Sample FR)	Weight	Weighted FR*Proportion
	Kept	20%	N/A	100%	0.0	0.0			100%	0.000
			Transferred	50%	0.5	0.2			74%	0.147
Participant	Discarded	80%	Destroyed	50%	1.0	0.4	0.60	0.03	74%	0.294
			Transferred	60%	0.5	0.2			26%	0.064
Nonparticipant	Discarded	N/A	Destroyed	40%	1.0	0.3	0.56	0.05	26%	0.085
Total Weighted										**0.589**

5.4 Induced Replacement Ratio (INDUCED_REPLACEMENT_RATIO)

Evaluators must account for replacement units *only* when a recycling program induces replacement (that is, when the participant would *not* have purchased the replacement refrigerator in absence of the recycling program). As previously noted, the purchase of a refrigerator in conjunction with program participation does not necessary indicate induced replacement. (The refrigerator market is continuously replacing older refrigerators with new units, independent of any programmatic effects.) However, if a customer would have not purchased the replacement unit (putting another appliance on the grid) in absence of the program, the net program savings should reflect this fact. This is, in effect, akin to negative spillover.

The most effective way to estimate the proportion of households induced to replace their appliance is through participant surveys. As an example, participants could be asked, “Would you have purchased your replacement refrigerator if the recycling program had not been offered?”

Since an incentive ranging from $35 to $50 may be insufficient to motivate the purchase of an otherwise unplanned replacement unit (which can cost in excess of $1,500), it is critical that evaluators include a follow-up question to confirm the participants’ assertions that the program alone caused them to replace their refrigerator. For instance, participants could be asked, “Let me be sure I understand correctly. Are you saying that you chose to purchase a new appliance because of the appliance recycling program, or are you saying that you would have purchased the new refrigerator regardless of the program?”

The proportion of total respondents indicating that the program induced them to purchase a replacement refrigerator then represents the induced replacement ratio.

DRAFT

5.5 Annual Energy Consumption of Replacement Refrigerator (UEC_REPLACEMENT)

Survey the participants to determine which recycled refrigerators were replaced. When replacement is indicated, participants should be asked if the replacement refrigerator was a new high-efficiency unit, a new standard-efficiency unit, or a used unit. Once this is known, this protocol recommends that evaluators estimate the REPLACEMENT_UEC using a weighted average of the replacement scenarios and the following UEC for each scenario:

- ***New High-Efficiency Unit***: Determine UEC using the ENERGY STAR® Refrigerator Savings Calculator[28] or a utility-specific or regional Technical Reference Manual[29]
- ***New Standard-Efficiency Unit***: Determine UEC using the ENERGY STAR Refrigerator Savings Calculator.
- ***Used Unit (of Unknown Efficiency)***: Determined UEC through the gross savings analysis.

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[30]: RUL = EUL/3
- RUL can be estimated using survival analysis (when appropriate data are available).[31]

---

28 www.energystar.gov/ia/business/bulk_purchasing/bpsavings_calc/Consumer_Residential_Refrig_Sav_Calc.xls?b3a0-c7cb

29 Such as that used by the Regional Technical Forum: http://www.nwcouncil.org/energy/rtf/measures/res/ResRefrigerators_v2_1.xlsm

30 This formula was obtained from the *Database for Energy Efficient Resources* (http://www.energy.ca.gov/deer/).

31 In an evaluation of the NV Energy appliance recycling program, ADM Associates used survival analysis using secondary data using 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.

6.2 Freezers

Although this protocol focuses on refrigerators, most utility appliance recycling programs also decommission stand-alone freezers. While differences exist between the evaluation approach for each appliance type (e.g., 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.

DRAFT

7 References

10 CFR 430.23(A1). 2008. (http://www.gpo.gov/fdsys/pkg/CFR-2011-title10-vol3/pdf/CFR-2011-title10-vol3-part430-subpartB-appA.pdf)

ADM Associates, Inc., Athens Research, Hiner & Partners, and Innovologie LLC. 2008. "Evaluation Study of the 2004-05 Statewide Residential Appliance Recycling Program."

The Cadmus Group, Inc. 2010. "Residential Retrofit High Impact Measure Evaluation Report." (http://www.calmac.org/publications/FinalResidentialRetroEvaluationReport_11.pdf).

Navigant Consulting. 2010. "Energy Efficiency/Demand Response Plan: Plan Year 2 (6/1/2009-5/31/2010) – Evaluation Report: Residential Appliance Recycling."

DRAFT

Whole-Building Retrofit Evaluation Protocol

1 Measure Description 1
2 Application Conditions of Protocol 1
3 Savings Calculations 3
3.1 General Approaches 3
3.1.1 Two-Stage 3
3.1.2 Pooled 3
4 Comparison Group Specification 5
4.1.1 Self-Selection and Freeridership 6
4.2 Recommendations by Program Characteristics 8
4.3 The Full-Year Specification 9
4.4 The Rolling Specification 9
4.5 The Two-Stage Approach 12
4.5.1 Stage 1. Individual Premise Analysis 12
4.5.2 Step 2.Applying the Stage 1 Model 17
4.5.3 Step 3.Calculating the Change in NAC 17
4.6 Stage 2. Cross-Sectional Analysis 17
4.6.1 Recommended Forms of Stage-Two Regression 17
4.6.2 Choosing the Stage-Two Regression Form 20
5 Pooled Fixed-Effects Approach 21
5.1 Recommended Form of Pooled Regression 21
5.1.1 Choice of Pooled Form 22
6 Measurement and Verification Plan 24
6.1 IPMVP Option C 24
6.2 Verification Process 24
6.3 Data Requirements and Collection Methods 24
6.3.1 Billing Data 24
6.3.2 Weather Data 27
6.3.3 Tracking Data 27
6.4 Analysis Dataset 28
6.4.1 Analysis Data Preparation 29
7 Sample Design 31
7.1 Program Evaluation Elements: Considerations for Other Program Types and Conditions 31
7.2 Alternative Comparison Group Specifications 31
8 References and Resources 33

Whole-Building Retrofit Evaluation Protocol

Ken Agnew, Mimi Goldberg, DNV KEMA

Whole-building retrofit programs focus on a building's energy performance overall. Typically, these programs involve installing a mix of energy-efficiency measures that, in combination, reduce the total energy consumption of a house or facility. Examples of whole-building retrofit programs are home weatherization, Home Performance with ENERGY STAR®, and many low-income programs. While whole-building retrofit programs generally target residential buildings, they may also target small commercial buildings.

1 Measure Description

Because whole-building retrofits involve the installation of multiple measures, the estimation of the total savings requires a comprehensive method for capturing the combined effect of the installed measures. The general method recommended for this type of program is a billing analysis—the analysis of consumption data from utility billing records. This method is consistent with the recommended International Performance Measurement and Verification Protocol[1] (IPMVP) Option C, Whole Facility. Option C is designed in part to address evaluation conditions that occur with a whole-house retrofit program.

The billing 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 billing analysis can be an effective evaluation technique for whole-house retrofit programs, and it addresses both how and when billing analysis should be used.

2 Application Conditions of Protocol

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 evaluation methods noted in this chapter are applicable when all of the following are true:

- The program offers a mix of measures affecting the whole building.
- The expected whole-building savings from the combination of measures supported by the program are expected to be of a magnitude that will produce statistically significant results given:
    - the natural variation in the consumption data,
    - the natural variation in the savings, and
    - the size of the evaluation sample.

[1] International Performance Measurement and Verification Protocol (IPMVP), which is considered the gold standard for evaluating energy-efficiency programs.

- The baseline for determining savings is the condition of the participating building before the retrofits were made, rather than the standard efficiency of the new equipment.
- There is sufficient consumption data available—in the form of monthly or bi-monthly utility billing records— for the participants[2].
- (Optional) Consumption data are available for the same timeframe as for the participants for one or more of the following groups: (1) previous participants—those who took part in the program before the timeframe of the current evaluation; (2) subsequent participants; or (3) those who are on a list for future participation in the program.

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 the Furnace Boiler protocol uses a billing analysis result and addresses the baseline issue described in the third bullet above.[3]

---

[2] Daily consumption data are now available from some billing systems. From the perspective of billing analysis evaluation, such data are a finer-grained form of the same basic data. The methods discussed here are primarily applicable to daily consumption data. There are issues unique to daily data, and one obvious concern is increased serial correlation in the modeling process and the resulting artificially low standard errors. (Note that this protocol also does not explore the additional opportunities that are available with the finer grain data.)

[3] As discussed under Section 7 of the Introduction chapter to the UMP Report, small utilities (as defined under the SBA regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

3 Savings Calculations

Because whole-house retrofit programs install multiple measures, the estimation of the total savings requires a comprehensive method for capturing the combined effect of all of the installed measures. The general approach recommended for this type of program is a billing analysis.

3.1 General Approaches

Two general billing analysis approaches are described here: "two-stage" and "pooled."

3.1.1 Two-Stage Approach

This approach is recommended in cases where there are: (1) a valid comparison group, and (2) sufficient consumption data for each building in the analysis. The Two-Stage method[4] consists of these activities:

- In ***Stage 1, the weather-normalized annual consumption (NAC) is estimated separately for each building*** in the analysis for both the pre- and post-program periods. The weather normalization for each building and period relies on a longitudinal regression analysis. Observations in these regressions correspond to usage over different bill intervals (typically, months) for the same building. For participants, the difference between the building's pre- and post-program NAC represents the program-related change in consumption plus exogenous change. For non-participants the pre-post difference represents only exogenous change.
- In ***Stage 2, a cross-sectional analysis is conducted on the Stage 1 output*** to isolate the aggregate program-related change from the observed changes in consumption. Depending on how the regression equation is specified, observations in the second-stage analysis are either the change in NAC for different customers, or the separate pre- and post-program year NACs for different customers and pre- and post- periods.

3.1.2 Pooled

The pooled approach combines all participants and time intervals into a single regression analysis. This is also referred to as a "time-series cross-sectional analysis" because its observations vary both across time and across individual buildings.

The pooled approach is appropriate under most scenarios described here, but it is particularly recommended when either of the following is true:

- There is not a valid, separate comparison group;
- The goal is to measure an average effect over multiple program years.

The conditions for obtaining reliable results in these situations are described in a later section under the heading "Pooled Fixed Fixed-Effects Approach."

For the evaluation of a whole-house retrofit program, the following are recommended:

---

[4] The two-stage billing analysis is not the same as the econometric "2-Stage Least Squares" regression method.

1. Use past and future (or “pipeline”) participants as the comparison group for the current program year. (See the details in the next section.)

2. Use a Two-Stage Approach unless the consumption data are too limited to produce good normalization models for individual buildings (as discussed below). In that case, use the pooled method.

3. Interpret savings carefully so they can be adjusted for freeridership as necessary. (Most billing analysis results are either gross savings or fall somewhere in between net and gross.) The following section discusses this issue.

The comparison group specification is described next, followed by the Two-Stage Approach using this comparison group. Then the pooled analysis using the same data is described.

4 Comparison Group Specification

Choosing the right comparison group is of central importance for a successful billing analysis. The goal of a billing analysis is to measure the change in building energy consumption from the pre-program period to the post-program period without including the effect of natural changes in consumption not due to the program. The comparison group makes it possible to remove these other changes in consumption—referred to here as exogenous changes—resulting from changes in fuel prices, general economic conditions, natural disasters, etc.[5]

The optimal evaluation scenario for a billing analysis is a randomized controlled trial (RCT) experimental design. This is essentially 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.

The control group sets the standard by which billing analysis comparison groups should be assessed. For an RCT, a sampling of eligible participants is randomly assigned to one of two groups before the program installations **(**treatment). This assures that the two groups—treatment and control—are probabilistically similar in every respect except for the offer of program treatment. The basic structure of this process is a "difference of differences." 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 represents the program-related change plus exogenous change.
- For the control group, the pre-post difference represents only exogenous change.

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. At present, in the context of energy-efficiency programs, true RCT is rare outside of certain types of behavioral programs.[6] The approach remains the gold standard, however, and provides a good illustration of the ideal characteristics of a control group.

Where a program is not designed as an RCT, 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 perform function as an experimental control group.

---

[5] While weather-related change is a form of exogenous change, it is controlled for in the models.

[6] There are multiple reasons why RCT has not been more widely employed. Until recently, evaluation concerns have been less likely to drive program planning. Also, RTC requires denying or delaying participation to a subset of the eligible, willing population and, under some approaches, it involves giving services to people who either do not want them or may not use them. The importance of RCT to the evaluation process is motivating program administrators to consider incorporating RCT into their program structures more frequently.

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. 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 non-participants.

In theory, many important characteristics can be controlled for; 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 freeridership. One of the many possible characteristics that could define a program participant is the intent to perform energy-efficiency activity regardless of program support. 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.

4.1.1 Self-Selection and Freeridership

The interaction between self-selection and freeridership 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 is freeriders.

In the RCT scenario, the control group does not have access to the program. The naturally occurring savings generated by this 5% in the control group is part of the pre-post non-program, exogenous change. The savings from this 5% of natural adopters in the control group will equal the savings for the 5% natural adopters in the treatment group. This naturally occurring portion of treatment-group savings will thus be cancelled out by the corresponding naturally occurring savings in the control group in the difference of difference calculation. That is, in a true RCT design, naturally occurring energy-efficiency savings—and, in the process, freeridership—are fully removed from the estimate of program-related savings. The result is a “net” estimate of savings, that is, program savings net of freeridership.

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 energy-efficiency 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 energy efficient-inclined population. Customers and contractors inclined toward energy efficiency have little reason not to take advantage of the rebates. This is likely to lead to an over-representation of natural adopters in the participant population, as compared to the general incidence in the population. This, then, affects in multiple ways the level of savings and freeridership that will be measured by the billing analysis.

- First, any comparison group developed after the fact from those who chose not to participate will tend to have a lower percentage of energy-efficient furnace installers (in this example) than would a true control group. To the extent that this is the case, the comparison group will not control for the full extent of natural energy-efficient furnace installations had the program not been in place.
- Second, the treatment group includes a higher proportion of natural energy efficient adopters than the general population, due to self-selection into the program. These households increase the freeridership rate beyond the natural level of natural adopters in the eligible population.
- Finally, the more general concerns regarding self-selection are still present. Because of their natural inclination to adopt energy efficient, the participants are likely to exhibit different energy-consumption characteristics than the general population.

These are the key factors that make it difficult to define fully the measured differences in consumption for the participant and comparison groups. 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 a mix of net and gross savings.

In the extreme, all household 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 and would not control for any natural energy efficient activity. This savings estimate would retain all of the freerider 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 as the participants and, as a result, are unlikely to exhibit any natural energy-efficient activity of the sort under evaluation. The use of customers who have participated in the same program in a recent year—or will participate in the near future (pipeline)—avoids most of the concerns related to self-selection bias. 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.

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. As a result, a comparison group created from previous and future participants may be as similar to current-year participants as is possible outside of an

RCT. Thus, the use of such a comparison group is likely to produce a gross estimate of savings that is unbiased due to self-selection.

4.2 Recommendations by Program Characteristics

The billing 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.

Table 1. Program Characteristics, Comparison Group Specifications, and Billing Analysis Structure and Interpretation

Program Condition	Billing Analysis Form	Comparison Group	Gross or Net Savings	Unknown Biases
1. Randomized Controlled Trial, Experimental Design	Two-Stage or Pooled	Randomly Selected Control Group	Net	Spillover, if it exists
2. Stable Program & Target Population Over Multiple Years	Two-Stage	Prior and Future Participants	Gross	Minimal
3. Participation staggered over at least one full year	Pooled	None: Pooled specification with Participants only	Gross	Minimal
4. Not randomized, not stable over multiple years, participants similar to general eligible population, nonparticipant spillover minimal	Two-Stage or Pooled	Matched comparison group	Likely between gross and net	Self-selection[7] and Spillover
5. Not randomized, not stable over multiple years, participants unlike general eligible population, nonparticipant spillover minimal	Two-Stage or Pooled	General Eligible Nonparticipants	Likely between gross and net	Self-selection and Spillover

Table rows 1, 2, and 3 provide at least one feasible approach for any whole-building retrofit program. Experimental design is still somewhat rare, but for many of the reasons discussed in this document, it is becoming more-widely used. A stable program makes possible the opportunity to obtain an unbiased estimate of savings using the Two-Stage approach.

Most other programs can be evaluated using the pooled approach. Rows 4 and 5 of the table list two relatively common approaches in the industry. These approaches produce an estimate that is a mix of net and gross savings. If this approach is used, then the result must be considered a conservative gross savings estimate with a known downward bias, to the extent freeriders still exist in the comparison group population. A separate freeridership analysis is required (for example, self-reported) to adjust all of these gross savings estimates to net savings estimates.

[7] The matched comparison should mitigate some self-selection to the extent that it is correlated with relative pre-period consumption, and this is an improvement over a non-matched, general population comparison group.

There are two ways to structure the analysis with past and future comparison groups: full year and rolling.

4.3 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 before the start of the current program year,[8] 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

4.4 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 utilizes 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.

8 It is counter-intuitive 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, similar in all ways to post-program participants. The difference-in-difference structure relies on an additive period-to-period change factor that works equally well with past or future participants.

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[9], 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; or
- 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

[9] This may indicate changes in the program or the program participants that may affect whether this is, in fact, a valid comparison group.

the analyst has the freedom to create these comparison group pre- and post- data periods using exactly the same distribution as the current year participant dates.

3.4 Basic Data Preparation

Before a billing analysis can be performed, the following activities must be done. The details of these steps are provided later in this section.

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 is 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.

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°F through 75°F—for each

day included in the analysis. (This activity is detailed in the section titled, "Data Requirements and Collection Methods.")

10. ***Calculate average daily degree days*** for the exact dates of each bill interval in the consumption data.

4.5 The Two-Stage Approach

4.5.1 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 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.[10] (The theory regarding the underlying structure is discussed in materials for and articles about the software.[11]) Stage 1 of the analysis can be conducted using PRISM or other statistical software.

4.5.1.1 Step 1.Fit the Basic Stage 1 Model

The degree-day regression for each premise and year (pre or post) is modeled as:

$$E_m = \mu + \beta_H H_m + \beta_C C_m + \varepsilon_m$$

where

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 over those dates;

---

[10] PRISM (Advance Version 1.0) Users' Guide.Fels, M.F., and k Kissock, M.A. Marean and C. Reynolds.Center for Energy and Environment Studies, Princeton New Jersey. January 1995.

[11] Energy and Buildings: Special Issue devoted to Measuring Energy Savings: The Scorekeeping Approach. Margaret F. Fels, ed. Volume 9 Numbers 1&2, February/May 1986.

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 over 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.

4.5.1.2 Stage 1 Model Selection

4.5.1.2.1 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 the savings estimation uses NAC only ***and*** the decomposition of usage into heating, cooling, and base components is not of interest. When 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 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.

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, adjusted R^2, AIC or BIC[12]) from among all of these models.

---

[12] 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.

4.5.1.2.2 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 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 me estimated for each of these options: a heating-cooling model (HC), heating only (HO), cooling only (CO), and no degree-day terms (mean value).

4.5.1.2.3 Degree-Days and Setpoints

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, and
- 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 setpoint and building dynamics, such as insulation, and internal and solar heat gains.

The average thermostat setpoints 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 (from55^{o}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.[13] (Note that the cooling degree-day base must always be higher than the heating degree-day base.)

---

[13] 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, H0 or C0), select as the final degree-day bases the values of τ_H, and τ_Ct hat 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.

4.5.1.3 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 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^oF for heating and 70^oF 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[14]:

- If the heating and cooling coefficients in the HC model have p-values[15]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.
    - 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.***

---

[14] Adjusted R2, AIC or BIC are also used.

[15] A measure of statistical significance.

4.5.2 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, $\beta_{H,}$ and β_C with the annual normal-year or typical meteorological year (TMY)[16] degree-days H_0 and C_0 calculated at the site-specific degree-day base(s), τ_H and τ_H. 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.

$$NAC = \mu * 365 + \beta_H H_0 + \beta_C C_0$$

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.

4.5.3 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.

4.6 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.

4.6.1 Recommended Forms of Stage-Two Regression

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 since they have different central tendencies and variances.

4.6.1.1 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.

$$\Delta NAC_j = \beta + \gamma I_j + \varepsilon_j$$

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

---

[16] Discussed in Section 4, Measurement and Verification Plan.

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.

4.6.1.2 Form B. Multiple Regression With Program Dummy Variables

This form allows for the estimation of savings for different measures. It may also include other available premise characteristics that can improve the extrapolation of billing analysis results to the full program population.

$$\Delta NAC_j = \Sigma_q \beta_q x_{qj} + \Sigma_k \gamma_k I_{kj} + \varepsilon_j$$

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; or

- With multiple dummy variables I_k and a single characteristics variable x (other than the intercept); or
- 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 a low-income program, these variables may be obtained from tracking data collected the same way across the program years.

Form C. Statistically Adjusted Engineering (SAE) 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.

$$\Delta NAC_j = \Sigma_q \beta_q x_{qj} + \Sigma_k \gamma_k I_{kj} + \Sigma_k \rho_k T_{kj} + \varepsilon_j$$

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, 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, or
- With multiple measure groups k and a single characteristics variable x (other than the intercept), or
- With 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, unless one of their associated coefficients is found to be statistically insignificant.

A simpler 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—or if it varies in ways that are not well correlated with actual savings—then the inclusion of the tracking estimate will not improve the fit. Thus, the dummy-variable version is preferred.

4.6.2 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, ***and***
- 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, ***and***
- 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, ***and***
- 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 billing 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.

5 Pooled Fixed-Effects Approach

The pooled approach addresses exogenous change without the inclusion of a separate comparison group. In this model, participants who received a measure installation during a certain time interval serve as a steady-state comparison for other participants in each other time interval.

Almost all observations include premises that are still in their pre-installation period ***and*** premises in that are in their post-installation period, so the effect of post- versus pre- is estimated to control for exogenous trends.

The basic structures of the site-level and the second-stage billing model are effectively combined in the pooled approach. All monthly participant 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) and average;
- Overall heating and cooling components; and
- A post-installation indicator variable capturing the change in the post-installation period.

5.1 Recommended Form of Pooled Regression

The recommended pooled model equation is as follows:

$$E_{im} = \mu_i + \phi_m + \Sigma_k \beta_{Hk} I_{kj} H_{jm} + \Sigma_k \gamma_k I_{kj} P_m + \Sigma_k \gamma_{Hk} I_{kj} H_{jm} P_m +$$

$$\Sigma_{kq} \beta_{Hkq} I_{kj} H_{jm} x_{qj} + \Sigma_{kq} \gamma_{kq} I_{kj} x_{qj} P_m + \Sigma_{kq} \gamma_{Hkq} I_{kj} H_{jm} x_{qj} P_m + \varepsilon_{im}$$

Where all variables have already been defined except for these:

μ_i	=	Unique intercept for each participant i,
ϕ_m	=	0/1 Indicator for each time interval *m*, time series component that track systematic change over time
P_m	=	0/1 Indicator variable for the post-installation period.

This specification only includes heating terms (H_{im}) for a gas analysis; however, analogous cooling terms should be included for an electric pooled model.

The parameter interactions that include the variable P_m capture the savings in the post-installation period. The inclusion of the read interval fixed-effects controls for exogenous factors specific to each month, and to first order eliminates the correlation across customers ij of residuals, ε_{im}, for a given month m.

If there is any intent to use the heating or cooling components of the model separately, the model should be fit across a range of degree-day base combinations. The highest R^2 is used to determine the optimal degree-day base combination[17].

From the fitted equation:

- The mean program savings must be calculated using the coefficients on all of the post-period dummy variable components, annual normal or TMY heating, and/or cooling degree-days for participants with measure k ***and*** the mean household characteristics (square footage, etc.) for households with measure k. That is, the estimated mean program savings for measure group k is

  $$S_k = \gamma_{k*}365.25 + \gamma_{Hk}H_{0k} + \Sigma_q \gamma_{kq}x_{qk_} + \Sigma_q \gamma_{Hkq}H_{0k}x_{qk_}$$

- WhereH_{0k} is normalized or TMY degree-days at the appropriate base for the subset of households with measure k, x_{qk} is the mean value of characteristics variable x_q for customers who received measure k.
- The coefficient ϕ_m is a monthly estimate of mean change or trend unrelated to the program. Because of the fixed-effects structure, these estimates represent the delta from the month or months left out of the model. That is, they are not mean zero and must be must be included if pre-treatment consumption is to be calculated.

In general, the increased complexity of the pooled approach requires additional care by the evaluator. The estimates of savings and consumption developed from any model must be carefully constructed and vetted against raw data. Developing a parallel two-stage model as a point of comparison for pooled model quality control should be considered.

5.1.1 Choice of Pooled Form

The pooled approach is recommended if:

- There is not a valid nonparticipant comparison group, or
- The goal is to measure an average savings effect over multiple program years.

In addition, the pooled approach requires both of the following:

- ***A balance of participant installation intervals across at least three billing intervals,*** preferably more. Having a balanced participation across three intervals would ensure that two-thirds of the participants provide a steady-state comparison during each interval of change. In the extreme, with only a single start date (as with a program that starts mailing comparative usage reports to homes at the same time), the model

---

[17] Note that the pooled model estimates average the heating and cooling degree day bases and average that slopes that are meant to represent the average across all homes in the model (or defined by interaction effects). This averaging can work well in many cases, but it can be difficult to determine when it may not work well. Therefore, if specific heating or cooling load components are of interest, the two-stage approach, which allows for house-specific degree day bases and heating/cooling slopes, may be a better choice.

fails to control for exogenous change across the change point. This explains the more stringent requirement for these programs of a randomly assigned experimental design.

- ***A balance of data between pre- and post-installation periods with respect to the number of data points per household and the seasonal coverage.*** Similar seasonal coverage in the pre- and post-installation is particularly important if measure savings are temperature sensitive. For gas heat modeling, the model should include at least one full winter in both the pre- and post- periods *and* some non-heating months. A full year of pre- and post-installation data removes concerns regarding imbalanced data.

The recommended specification includes the characteristics variables (x_j) for each house because of the importance of these factors:

- Having additional data to inform the overall average heating and cooling trends, and
- The changes in those trends due to the program.

In particular, it is useful to include a consistent square-footage variable. These characteristics data help compensate for the pooled approach's inherent lack of flexibility with respect to heating and cooling dynamics, as compared to the site-level model approach.

6 Measurement and Verification Plan

6.1 IPMVP Option C

The recommended IPMVP method is Option C (Whole Facility), which was designed in part to address evaluation conditions that occur with a whole-house retrofit program. The key reasons for using this method are these:

- The goal of the program is improvement of whole-house performance;
- Because multiple different measures are installed, the individual savings of each cannot be easily isolated because of interactive effects; and
- The expected savings are large enough to be discernible over natural variation in the consumption data, at least across the aggregate of program participants.

Major non-program changes in energy consumption are either not expected or will be adequately controlled for in the analysis.

6.2 Verification Process

This does not apply for whole house retrofit savings based on billing analysis.

6.3 Data Requirements and Collection Methods

A billing analysis requires data from multiple sources:

- Consumption data, generally from a utility billing system,
- Program tracking data, and
- Weather data.

This section describes the required data for a whole-house retrofit billing analysis and the steps for using these data correctly.

6.3.1 Billing Data

The consumption data used in a billing analysis are generally stored as part of the utility billing system. Since 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 billing 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 billing regression.

The remaining requested variables serve one of three purposes:

- Linking the billing data with other essential data sources (such as program tracking data and weather data);
- Providing information that facilitates the cleaning of the consumption data; or
- Providing data for characterizing the household so as to improve the quality of the regression models.

6.3.1.1 Billing 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 99th 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; however, do this 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 billing 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).

6.3.2 Weather Data

Weather data are used in the billing 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[18].

Use either primary 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 so as to produce evaluation results that are consistent with other studies within the utility. Many utilities are choosing to use normals constructed from fewer than 30 years, as are the standard NOAA norms.

A billing 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 that the same site. 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.

6.3.2.1 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 concern regarding a weather data source, consider using a more distance 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.

6.3.3 Tracking Data

The program tracking data provide the participant population, the pre- and post-installation time periods, and the number and type of measures for which savings are claimed. Frequently, the

---

[18] The National Oceanic and Atmospheric Administration (NOAA) produces 30-year normal weather series composed of average temperature for each hour over the time period. These normals are updated every decade. National Renewable Energy Laboratory produces typical meteorological year (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.

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 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 year may be used to define a control group for a Two-Stage analysis.

6.4 Analysis Dataset

Using the account numbers in the two datasets, the final analysis dataset combines the tracking data and the billing 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 of the unique read intervals, some evaluators resort to calendarized data.[19] Except in special cases, calendarization should not be

[19] Calendar month consumption is estimated as a weighted average of the bill readings that cover that month.

used for this kind of analysis because it undermines the direct matching between consumption and degree-days that is the basis of billing analysis. Multiple meter and multifamily analyses are examples of situation where calendarization may be the only way to aggregate data series on difference schedules.

6.4.1 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 billing 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 billing 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.

    - 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 billing 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.

7 Sample Design

Sample design is generally not required for whole-house retrofit billing analyses because this type of evaluation is performed on the full, relevant program population.

7.1 Program Evaluation Elements: Considerations for Other Program Types and Conditions

The methods described above are used in whole-building program evaluation for an ongoing, stable residential program. Similar methods can be used for the following: (a) other whole-premise programs for the residential population; (b) whole-premise programs for small commercial populations; and (c) with modification, for new construction.[20] Whole-premise billing analysis is also used for other types of programs, such as single-measure rebate programs and recycling programs.[21] In this section, we discuss the alternative comparison group specification to use in these situations.

7.2 Alternative Comparison Group Specifications

In some cases, it is not practical to use past or future participants as a comparison group, or to conduct a pooled billing 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, as discussed above.
- 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, a two-stage 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.

Matching is accomplished by: (1) Determining the mix in the participant population, and (2) selecting a stratified nonparticipant sample with the corresponding mix from those customers

---

[20] Discussed in a separate chapter.

[21] See Chapter on Furnaces and Boilers

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); and
- Energy end uses.

8 References and Resources

ASHRAE recommends that the following ASHRAE documents be listed as either a reference or resource document.

- Performance Measurement Protocols for Commercial Buildings, ASHRAE, 2010
- Research Project 1050 Development of a Toolkit for Calculating Linear, Change-Point Linear and Multiple-Linear Inverse Building Energy Analysis Models, ASHRAE Research Report, 2004.
- ASHRAE Guideline 14-2002 Measurement of Energy and Demand Savings, ASHRAE, 2002.
- ASHRAE Guideline 14-2002R (Revision of Guideline 14, currently in process, publication date TBD).

Metering Protocols

Dan Mort, ADM Associates, Inc.

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 conservation measures (ECMs).

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 ECM. This chapter describes the physical properties measured in the process of evaluating ECMs 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 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.”)

1.1 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 ECM 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 ECM, refer to the Sample Design Protocol chapter to determine how many units need to be metered.

1.2 Ensuring Precision and Verification

The accuracy of a measurement is typically proportional to the cost of the instrument and 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 the NIST[1] 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 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

[1] National Institute of Standards, website: nist.time.gov

section describes these properties and their respective measuring methodologies. The corresponding equipment descriptions are included in a subsequent section.

2.1 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.[2]

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)[3], and power factor (pf). Although direct current (dc) voltage is used to power some types of equipment, utility transmission to customers occurs in the form of alternating current (ac). For this discussion, A and V are expressed in terms of alternating current, and the values measured or recorded are the root mean squared (rms) values. In general terms, rms is the common presentation of ac 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.

2.1.1 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.

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, or

[2] Note that power metering is also referred to as kW metering.
[3] Current metering is also referred to as Amp metering.

- 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.

2.1.2 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.

2.1.3 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 meters' 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.

2.2 Temperature

Temperature is an indirect parameter that is incorporated into the calculation of energy use or estimated savings for some types of ECMs. 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.

2.2.1 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

2.2.2 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.

2.2.3 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.)

2.2.4 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.

2.3 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 those same conditions.

Humidity is measured when estimating the enthalpy or energy content of air in an 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.

2.4 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).

2.4.1 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.

2.4.2 Duct Air Flow

Air flow measurements are most commonly needed for ducts carrying conditioned air, and these measurements can be made by anyone trained in the technique. Note that air or gas flow rates should be normalized to standard temperature and pressure conditions (68°F and 14.7 psi).

The preferred methods for measuring air flow 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; or
- 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.

Air flow 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.

2.4.3 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.

2.5 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 (VFD) is installed.

2.5.1 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.

2.5.2 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 air flow 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.

2.6 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 ECM 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 ECM is installed, the illuminance measurements justify applying additional savings. There are no specific qualifications required for personnel conducting illuminance measurements.

2.6.1 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 ECM lighting is on and again when it is off. The difference in the two values is the illuminance attributable to the ECM lighting.

2.7 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.

2.8 Normalizing Conditions

In many cases, to normalize the energy use of the ECM, 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 (TMY) 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

3 Levels of Measurement

Electric loads should be metered at the level appropriate for the type of ECM. 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); and

- The utility meter itself.

3.1 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) in the disconnect box at the motor; or (3) in 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 of 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 are at the MCC.

3.2 Aggregation of Like Loads

Since lighting is generally updated as a retrofit throughout a wide area of a facility or throughout an entire facility, metering a representative sample usually suffices, rather than conducting metering for a census of fixtures.

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 the Sample Design Protocol chapter.)

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.

3.3 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.

4 Duration of Measurement & 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 ECM and magnitude of the load.

4.1 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.

4.2 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 M&V effort. Cooling loads, for example, should be monitored during the hottest part of the year.

4.3 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), or
- Operate on planned schedules (such as educational facilities)

4.4 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[4]. 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 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:

- By the type of measurement. (This equipment can be sub-categorized by dedicated single measurement or by multi-purpose and multi-channel.)
- By 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.

5.1 Electrical

Electrical measurement equipment can be categorized as:

[4] Time stamp resolution is generally one second.

- handheld (or portable) power meters
- watt-hour transducers
- meter recorders
- current transformers (CT)

5.1.1 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.

5.1.2 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.

5.1.3 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).

5.1.4 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.

5.1.4.1 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.

5.2 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.

5.2.1 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.

5.2.2 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.

5.2.3 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.

5.3 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 & 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:

- The 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.

5.3.1 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.

5.3.2 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.

5.3.3 Differentials

Measurements used to estimate differential temperature—such as supply and return air—should use a matched pairs of sensors.

5.4 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.

5.5 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.

5.5.1 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.

5.5.2 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.

5.5.3 Instantaneous

Use digital pressure gauges for conducting instantaneous readings.

5.6 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.

5.6.1 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.

5.6.2 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.

5.6.3 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.

5.7 Other Sensors

Other commonly used sensors and meters are these:

- Occupancy sensors,
- CO_2 sensors,
- Combustion gas analyzers
- Solar radiation sensors (such as a pyranometers)
- Wind speed sensors
- BTU meters

When selecting the desired level of accuracy for each of sensor types, consider both cost-effectiveness and the importance of the measurement to the final savings estimations.

5.7.1 Digital Cameras

Digital cameras are very useful in documenting metering equipment before and after its installation and during the evaluation of the ECM.

5.8 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.

5.8.1 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.

6 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.

6.1 Data Storage

While the memory storage capacity of data loggers varies widely, ideally, loggers 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.

6.2 Retrieval

Evaluation of ECMs 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 land lines or cell phones, Ethernet and Internet, and other digital contact via local networks.

6.3 Handling

After retrieving the data, make backup copies immediately. For the data files, use a filename convention that includes the site, ECM, 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.

7 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 ECM to be evaluated.

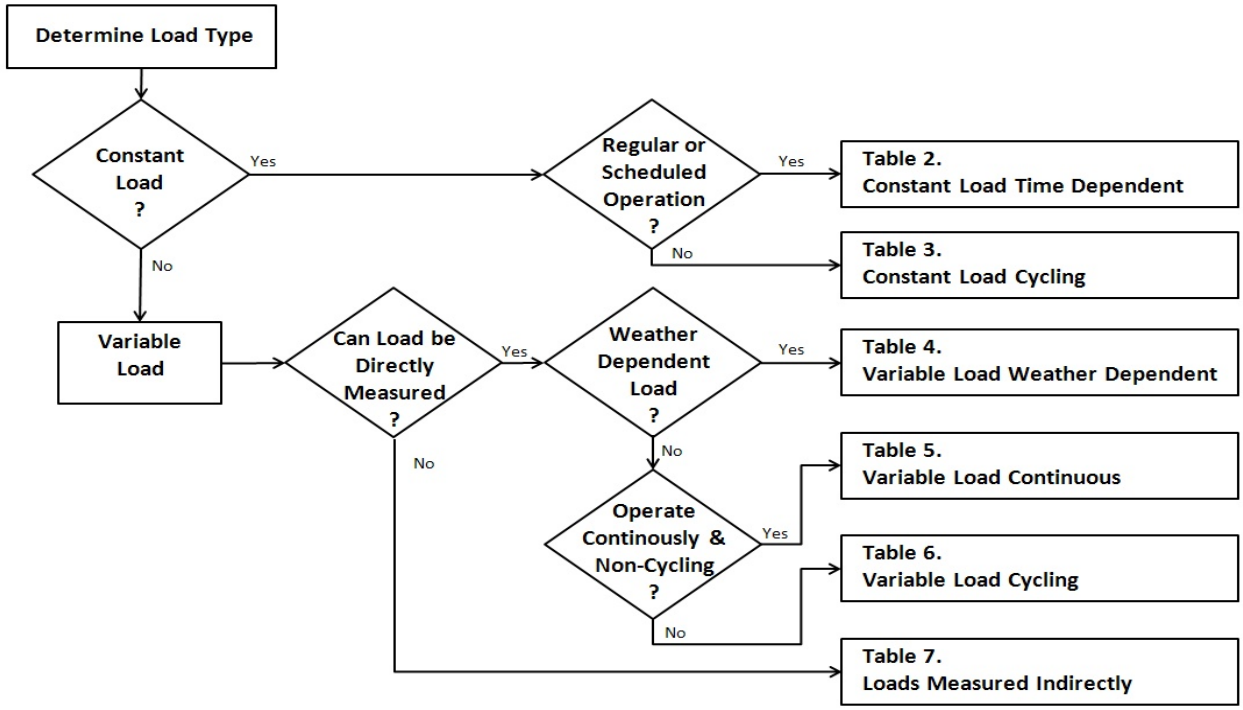
Some measures (such as building envelopes) do not directly use energy, but they impact energy use. In those cases, the end use that would be metered is the energy-using equipment impacted by the measure. In general, these categories are listed in increasing order of metering complexity. The example end-uses 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, since 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.
Weather-Dependent Variable Load	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



7.1 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 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, or
- 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.

7.2 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 or bi-level controls Refrigerators & freezers Water heaters, electric Plug-in loads Household & office electronics ECM fans	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

Example End-Use	Rigor Level
Electric ovens or grills	Additional measurement: Instantaneous Volts, Amps, kW, and power factor. Duration: Two weeks Interval: 2 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. Weather-Dependent Variable Load

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: 2 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: 2 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 24Vac coil or combustion air fan motor. Additional measurement: Measure burner flow rate using utility meter with all other loads off & 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 & 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*pf. Assume V and pf. Duration: One month Interval: 15 minutes

8 References and Resources

Alereza, T., M. Martinez, D. Mort, Monitoring of Electrical End-Use Loads in Commercial Buildings, ASHRAE Transactions 1989 V. 95, Part 2.

American Society of Heating Refrigeration and Air Conditioning Engineers (ASHRAE). 2010. Performance Measurement Protocols for Commercial Buildings.

ASHRAE Guideline 14-2002, June 2002, Measurement of Energy and Demand Savings.

ASHRAE. 1996. Research Project 827 Methodology Development to Measure In-situ Chiller, Fan, and Pump Performance, ASHRAE Research Report.

ASHRAE Handbook, 1995 HVAC Applications, Chapter 37: Building Energy Monitoring

Bonneville Power Administration, September 2010, End-Use Metering Absent Baseline Measurement: An M&V Protocol Application Guide, Version 1.0, Available at: http://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. Available at: http://www.calmac.org/publications/California_Evaluation_Framework_June_2004.pdf

Federal Energy Management Program (FEMP). August 2011, Best Metering Practices Release 2.0. Available at: http://www1.eere.energy.gov/femp/pdfs/mbpg.pdf

Federal Energy Management Program (FEMP). September 2000. Measurement & Verification Guidelines for Federal Energy Projects Version 2.2. DOE/GO-102000-0960. Available at: http://www.nrel.gov/docs/fy00osti/26265.pdf

Federal Energy Management Program (FEMP). April 2008. Measurement & Verification Guidelines for Federal Energy Projects Version 3. Available at: http://www1.eere.energy.gov/femp/pdfs/mv_guidelines.pdf

Institute of Electrical and Electronics Engineers (IEEE), ANSI/IEEE Std. 120-1989. 1989. Master Test Guide for Electrical Measurements in Power Circuits.

IPMVP. 2010. International Performance Measurement and Verification Protocol Volume 1: Concepts and Options for Determining Energy and Water Savings. EVO 10000 – 1:2010. Washington, D.C.: Efficiency Valuation Organization. Available at: http://www.evo-world.org/index.php?option=com_form&form_id=38.

Metering Your End-Use Needs, Conference Proceedings June 1-3, 1992, Bend OR

NEC. National Electric Code 2011. NFPA 70

NIOSH. July 1999. Preventing Worker Deaths from Uncontrolled Release of Electrical, Mechanical, and Other Types of Hazardous Energy. Publication No. 99–110. Available at: http://www.cdc.gov/niosh/docs/99-110/pdfs/99-110.pdf

OSHA. Occupational Saftey and Health Administration. Standard 1910 emphasis on Subpart S – Electrical. Available at: http://www.osha.gov/pls/oshaweb/owasrch.search_form?p_doc_type=STANDARDS&p_toc_level=1&p_keyvalue=1910

Peak Demand and Time-Differentiated Energy Savings Cross-Cutting Protocol

Frank Stern, Navigant

Energy-efficiency savings are often expressed in terms of annual energy, presented as kWh/year. However, for a full assessment of the value of these savings, it is usually necessary to consider peak demand and time-differentiated savings.

1 Purpose of Peak Demand and Time-differentiated Energy Savings

Energy efficiency may reduce peak demand and, consequently, the need for investment in new generation, transmission, and distribution systems. This reduction in the need for new investment —also called "avoided capacity costs"—has value, and to estimate this value, 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 be a substantial portion of the value of an energy-efficiency measure, particularly for measures that produce savings coincident with the system peak. The need to estimate peak demand savings is becoming more important as regional transmission organizations (RTOs, such as PJM and ISO-New England) allow energy-efficiency resources to bid into the forward capacity markets and earn revenues.[1]

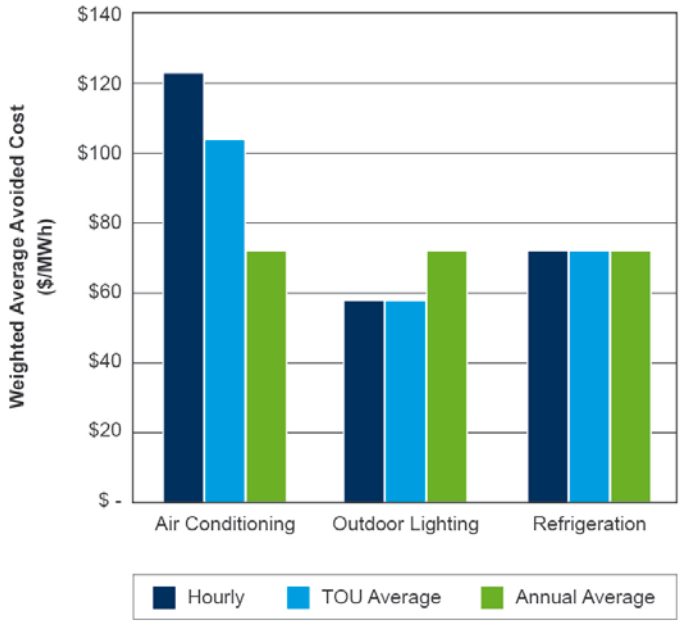
In addition to considering peak demand savings, evaluators often must calculate time-differentiated energy savings. This is because avoided energy costs are typically provided in terms of costing periods. These costing periods divide the 8760 hours of the year into periods with similar avoided energy 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).[2]

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 reduced energy mostly at night is not as valuable as one that reduced energy mostly during summer afternoons, as shown in Figure 1.

[1] These are where the regional transmission markets obtain the resources for ensuring system reliability. Providers of energy efficiency can bid into these markets on an equivalent basis to supply-side resources. Bids must be supported by measurement and verification.

[2] Avoided energy costs tend to be higher during periods of higher demand because generating units available during those times tend to have lower efficiency and higher operating costs.

Figure 1. Consideration of Time-Differentiation in Energy Savings Significantly Affects Estimates of the Value Savings



Source: (U.S. Environmental Protection Agency and Department of Energy 2006)

As another example, air-conditioning efficiency has higher value when hourly savings and costs are considered, because usage is higher when avoided costs are higher. Outdoor lighting, however, has lower values when hourly savings and costs are considered, because that usage is typically off-peak.

Peak demand and time-differentiated energy impacts are more difficult to measure than annual energy savings impacts (York 2007), so additional metering or simulation analysis may be needed to estimate these impacts accurately. Peak demand savings and time-differentiated energy savings can be estimated with:

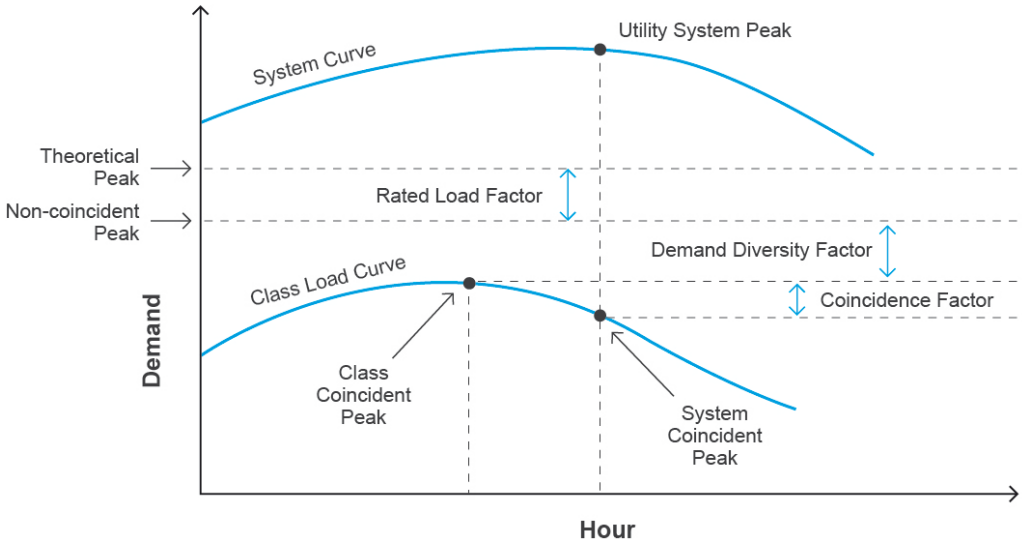
- Engineering algorithms,
- Hourly building simulation modeling,
- Interval meter data analysis,
- End-use metered data analysis,
- Survey data on hour of use, or
- Combined approaches.

Peak savings are estimated over a peak period This period can range from one hour per year to several hours per day during a season.

2 Key Concepts

Understanding demand savings requires understanding the relationship between several factors, as shown in Figure 2.

Figure 2. Demand Savings Relationships



Source: (Jacobs 1993)

Note: Rated load factor, demand diversity factor, and coincidence factor are sometimes combined and referred to as "coincidence factor."

These brief definitions describe the key factors.

- **Peak period.** The period during which peak demand savings are estimated. (As previously noted, this period can range from one hour per year to several hours per day during a season.) Some utilities have a winter and summer peak period.
- **Theoretical peak** is the usage of a population of equipment if all were operating 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 are not all operating at the same time across all buildings throughout the region. For example, if a maximum of 7 of 10 installed CFLs are on at any given time, then the diversity factor is 0.7.

- **Coincidence factor** is 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 non-coincident peak demand. The peak demand use for a given building and end use are typically not aligned exactly with the utility system peak, which is how the 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.

  Some technical references use the term "coincidence factor" to mean the product of rated load factor, demand diversity factor, and coincidence factor. NEEP defines it as, "The ratio of the average hourly demand during a specified period of time of a group of electrical appliances or consumers to the sum of their individual maximum demands (or connected loads) within the same period." (Northeast Energy Efficiency Partnership 2011).

The following terms are also are important to understanding the concepts of 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 8760 hours/year = 8760 MWh
- **Load factor.** The ratio of average energy savings to peak energy savings. This is also known as "peak coincidence factor" (NYSERDA 2008). More generally, load factor is the average demand divided by any number of peak demands, such as load factor at the time of system peak and load factor at the time of non-coincident peak.

$$Conservation\ Load\ factor = \frac{Energy\ savings}{Peak\ demand\ savings \times 8760\ hours}$$

- **Loss of load probability (LOLP).** The likelihood that a system will be unable to meet demand requirements during a period. LOLP can be used to distribute avoided capacity costs to each hour of the year.

3 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).

Approaches can also be combined to leverage available information. Some approaches for estimating annual energy savings (such as monthly billing data analysis) do not provide peak demand savings directly. However, these other approaches can be used with load shapes for analyzing peak impact.

3.1 Engineering Algorithms

Peak demand savings can be estimated using algorithms, as shown in Equation 1. This equation is similar to those used for energy savings (shown in Equation 2), except that the demand equation has diversity factor and coincidence factor in place of the full load hours.

Equation 1. Basic Demand Savings Equation

$$\Delta kW_{gross} = units \times RLF \times \left[\left(\frac{kW}{unit}\right)_{base} - \left(\frac{kW}{unit}\right)_{ee}\right] \times DF \times CF \times (1 + HVAC_d)$$

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. Basic Energy Savings Equation

$$\Delta kWh_{gross} = units \times RLF \times \left[\left(\frac{kW}{unit}\right)_{base} - \left(\frac{kW}{unit}\right)_{ee}\right] \times FLH \times (1 + HVAC_c)$$

Source: (TecMarket Works 2004)

3.2 Hourly Building Simulation Modeling

Hourly building simulation modeling (IPMVP[3] Option D) can produce hourly savings estimates for whole buildings as well as for specific end uses. Consequently, 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.

Simulation models are most applicable for HVAC, shell measures, and the interactive effects of HVAC with other measures. Simulation modeling requires an experienced modeler with an understanding of energy engineering. Hundreds of building energy simulation programs have been developed over the past 50 years (Crawley 2005).

Note that using this method does not necessarily provide an estimate of diversified demand. If a single, typical building is used, demand savings would be overstated due to lack of consideration of diversity, which tends to smooth out spikes in usage seen in individual buildings. Consideration of diversity requires either using average schedules or simulating a sample of buildings with different sizes, climate, and schedules.

[3] International Performance Measurement and Verification Protocol (IPMVP), which is considered the gold standard for evaluating energy-efficiency programs.

3.3 Billing Data Analysis

Billing data analysis (IPMVP Option C) can be used to develop monthly estimates of savings. (Billing analysis is discussed in the Whole-Building Retrofit Evaluation Protocol section.) This type of analysis entails statistical comparison of pre- and post-participation and/or participant and nonparticipant billing data to estimate savings. Complex statistical analysis may be required to control for non-programmatic influences, such as weather and economic conditions. Also, isolating the impacts of a specific measure can be difficult because the meter measures usage for an entire building.

Although the coincident peak is usually not reported, billing analysis is useful in estimating non-coincident peak demand when the data include monthly building peak demand for each costing period. In addition, billing data analysis can be used to derive a realization rate on an engineering algorithm for energy savings that may also be applied to a demand savings algorithm.

3.4 Interval Metered Data Analysis

Utility revenue interval meters can measure usage at in increments of 15 minutes or less. Since 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 the impacts of a specific measure can be difficult, and statistical analysis may be required to control for non-programmatic influences. With the advent of advanced metering infrastructure and the availability of obtaining hourly information, there may be additional statistical approaches (such as conditional demand type analysis on hourly data) that could be used to help develop estimates of demand savings.

3.5 End-use Metered Data Analysis

End-use metering data analysis (IPMVP Option A and Option B) can be an excellent 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 non-participant billing data. However, end-use metering eliminates most—if not all—of the difficulty of isolating the impacts of specific measures.

There are several cautions to consider:

- Savings should be normalized for weather and other confounding factors.
- Pre-installation meter data is difficult to obtain because of the logistics entailed in coordinating with customers. Without pre-installation data, baseline conditions must be estimated with engineering algorithms.
- End-use metering is costly, so it should be conducted strategically.
- An impact load shape may be different than a post-participation load shape. For example, lighting control impact shapes are different from the shape of the controlled lighting. (End uses have shapes with and without the efficiency measures in place and the difference is the impact shape.) Determination of some energy-efficiency shapes may require either pre-installation metering or reconstruction of the baseline shape.

- Sampling must be done carefully—see the Sampling Cross-cutting Protocol.
- The evaluator must consider the period over which to meter. How much time is required? Is a certain time, such as summer, critical?

The American Society of Heating, Refrigerating, and Air-Conditioning Engineers has developed a methodology to derive the diversity factors and provide the typical load shapes of lighting and receptacle loads for office buildings using end-use metered data (Abushakra 2001).

3.6 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 compact fluorescents are used during the summer from 3:00 p.m. to 6:00 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%.

Survey sampling should be done in conjunction with the techniques described in the Sampling Cross-Cutting Protocol section. However, relying on customer perception may result in significant inaccuracy.

3.7 Combined Approaches

Applying a combination of approaches facilitates using data from several sources to provide the best estimates of demand savings. For example, for a low-income program, billing data may be the best approach for estimating energy savings. Engineering algorithms can be used to develop energy and demand savings for each participant, and these participant energy savings can be the independent variables in a statistically adjusted engineering (SAE) billing analysis. (See the protocol titled Whole House Retrofit.) 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 is used to calibrate telephone 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 telephone survey produced an estimate of 5.0% for the metered sample and 5.5% for the entire telephone sample. The ratio of 6.1% to 5.0% would be applied to the 5.5% telephone sample estimate, resulting in an adjusted factor of 6.7%.

4 Summary of Approaches

Table 1 presents a summary of 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
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	Moderate	Moderate	Typically not useful for peak demand or on/off peak energy analysis.
Interval Meter Data Analysis	Moderate	High	Interval meter data not available for many customers. Becoming more feasible with proliferation of AMI.
End-Use Metered Data Analysis	High	High	Requires careful sampling and consideration of period to be metered.

5 Secondary Sources

Because of budget or time constraints, evaluators may choose to rely on secondary sources, rather than on the primary sources listed above.

5.1 Technical Reference Manuals

A technical reference manual (TRM) specifies savings or protocols for common energy-efficiency measures. A TRM is not a method for estimating savings, but a source of estimates or methods. Typically, TRMs provide deemed savings values that represent approved estimates of energy and demand savings. These savings are based on a regional average for the population of participants; however, they are not savings for a particular installation.

Although TRMs often provide industry-accepted algorithms for calculating savings, users should not assume that because an algorithm has been used elsewhere it is correct. Mistakes are common and should be expected.

5.2 Application of Standard Load Shapes

By applying load shapes to allocate energy consumption into costing period, peak demand and time-differentiated energy savings can also be estimated from energy impacts. A key resource of load shape data is the California Database for Energy Efficiency Resources (California Public Utility Commission 2011). These shapes may be derived from metering or simulation. The evaluator must consider the applicability of the shapes when climate-sensitive end uses are involved.

6 References and Resources

Abushakra, B., A. Sreshthaputra, J. Haberl, C. Clairidge. *Compilation of Diversity Factors and Schedules for Energy and Cooling Load Calculations.* ASHRAE, 2001.

California Public Utility Commission. "Database on Energy Efficient Resources." 2011.

Crawley, D., J. Hand, M. Kummert, B. Griffith. *Contrasting the Capabilities of Building Energy Performance Simulation Programs.* U.S. Department of Energy, 2005.

Fels, M. and K. Keating. *Measurement of Energy Savings from Demand Side Management Programs in U.S. Electric Utilities.* Annual Review of Energy and the Environment, 1993.

Jacobs, P., M. Arney, M. Keneipp, and M. Eldridge. *Engineering Methods for Estimating the Impacts of Demand-Side Management Programs.* Pleasant Hill, CA: Electric Power Research Institute, 1993.

Keating, M. Fells and K. *Measurement of Energy Savings from Demand Side Managemenet Programs in U.S. Electric Utiites.* Annual Review of Energy and the Environment, 1993.

Northeast Energy Efficiency Partnership. "Glossary of Terms: A Project of the Regional EM&V Forum." 2011.

NYSERDA. "Energy Efficiency Portfolio Standard Administrator Proposal." 2008.

TecMarket Works. *California Evaluation Framework.* Sacremento: California Public Utilties Commission, 2004.

U.S. Environmental Protection Agency, and Department of Energy. "National Action Plan for Energy Efficiency." Washington, DC, 2006.

York, D, M. Kushler, P. Witte. *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, 2007.

Sample Design Cross-Cutting Protocol

M. Sami Khawaja, Josh Rushton, Josh Keeling, The Cadmus Group, Inc.

1 Introduction 1
1.1 Chapter Organization 1
2 Overview 3
2.1 Sampling and Sample Design 3
2.2 Uncertainty and Efficiency 3
2.3 Confidence and Precision 5
3 Complex Evaluations: Designing for Multiple Objectives 8
4 Worked Examples 12
4.1 Measure- and Site-Level Evaluation Planning 12
4.1.1 Telephone Surveys 12
4.1.2 Verification Site Visits 15
4.1.3 End-Use Metering 18
4.2 Domain-Level Evaluation Planning 19
4.3 Portfolio-Level Evaluation Planning 21
5 Additional Considerations 24
5.1 Threats to Validity 24
5.2 Cost Considerations 24
5.3 Varying Uncertainty 26
5.4 Outcome of Interest 26
6 Appendix A. Sources and Types of Error 28
6.1 Sources of Uncertainty 28
6.2 Sources of Systematic Error 28
6.3 Sources of Random Error 30
6.4 Mitigating Systematic Error 31
6.4.1 Measurement Error 32
6.4.2 Non-Coverage and Non-Response 32
7 Appendix B. Fundamental Estimates and Uncertainty Calculations 34
7.1 Estimating a Population Proportion 34
7.2 Using a Sample Mean to Estimate a Population Mean 37
7.3 Using a Ratio Estimator to Estimate a Population Mean 38
7.4 Estimating a Difference or Sum 44

7.5 Estimating a Product 45
7.6 Summary of Analytical Techniques 46
8 Appendix C. Sample Design and Weighted Estimates 47
8.1 Simple Random Sampling 47
8.1.1 Sample Means with FPC 47
8.1.2 Population Proportions and Ratio Estimators with FPC 50
8.2 Stratified Random Sampling 53
8.2.1 Stratified Means 54
8.3 Stratified Proportions 56
8.3.1 Stratified Ratio Estimators 57
8.3.2 Summary of Estimators for Stratified Samples 58
8.4 Planning and Optimizing Stratified Designs 59
8.5 General Probability Samples and PPS 59
8.6 Two-Stage Sampling for Large Projects 61
8.7 Two-Phase (Nested) Sampling 64

DRAFT

1 Introduction

Evaluating an energy-efficiency program requires assessing the total energy and demand saved through all of the 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), since 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 the survey design chapter and in this chapter's Appendix A.

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.

DRAFT

- 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 principals 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.

DRAFT

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* savings, *ex post* 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, *ex ante* savings) for every member of the population. Other variables (especially *ex post* 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**.[1]

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 *ex ante* 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,[2] 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.

2.2 Uncertainty and Efficiency

Sample design is typically approached with one of two goals:

---

[1] 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 M&V. In this case, auxiliary data are known only for the larger sample, but not the entire population.

[2] Here, n is the sample size, and N is the population size.

1. ***To minimize estimator uncertainty, given a fixed amount of study resources.*** In this case, time and budget are the primary constraint. 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, and
- 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 *ex ante* 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

DRAFT

among these values.[3] 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 sample of 100 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.[4] 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 instance, a **confidence interval** (CI). A confidence interval 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 kWh, and the relative precision level is 10%, then the interval is 530 ±53 kWh).[5] 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 trade-off between precision and confidence.[6] As a result, any statement of precision without a corresponding confidence level is incomplete and impossible to interpret. For example, assume the average

---

[3] 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.

[4] This formulation ignores the finite population correction (see "Sample Means with FPC" in Appendix C).

[5] Note the counter-intuitive 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."

[6] 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.

DRAFT

savings among participants in an ENERGY STAR® appliance program is estimated as 1,000 kWh per year, and 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.[7] To calculate the bounds for such an estimator, first multiply the estimator's standard error by a *z*-value.[8] 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:[9]

$$\hat{x} - z \cdot \widehat{\mathrm{SE}}(\hat{x}) \leq x \leq \hat{x} + z \cdot \widehat{\mathrm{SE}}(\hat{x})$$

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.[10]

---

7 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.

8 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.

9 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, since 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.

DRAFT

Example 1-1
If a program's average savings is 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:

$$\begin{aligned} 10.31\text{ kWh} - 1.645 \cdot 1.70\text{ kWh} &\leq \quad \text{average savings} \\ &\leq \quad 10.31\text{ kWh} + 1.645 \cdot 1.70\text{ kWh} \end{aligned}$$

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.

DRAFT

[10] 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 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; or
- 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 and Appendix C.

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); and
- 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; and
- 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 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 (e.g., 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.[11]

   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 since 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.[12]

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 (e.g., 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		

[11] 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.

[12] The general principals provided in the appendices remain valid for alternative approaches, but we do not provide step-by-step guidance for all possible approaches.

DRAFT

4. ***Determine the data requirements and estimation strategies within each domain.*** For each group (e.g., prescriptive commercial program) or subgroup (e.g., offices), use the program database to identify important measure categories (e.g., lighting). Then for each measure category, determine estimation procedures and data needs based on the prevailing measurement and verification (M&V) protocol.

5. ***Record ex ante contribution, ex post uncertainty, and M&V costs for each stratum.*** For each stratum within a domain, determine total *ex ante* 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 ER) within each category.[13] Measures contributing significantly to total savings and exhibiting significant variability will receive highest levels of evaluation resources.[14] This will reduce the standard error and improve confidence intervals.

6. ***Estimate sample sizes within each domain.*** In the most straight-forward cases, the previous step will yield reliable cost, uncertainty, and *ex ante* total estimates. In such a case, implement the cost-weighted Neyman formula (Appendix C) 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 straight-forward Neyman allocation. In such cases, the planning process may be simplified by prioritizing measure categories with high *ex ante* 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.

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 evaluation dollar and to increase evaluation resources for these categories until higher-level precision estimates meet the evaluation targets.

---

[13] This may be based on previous studies' estimates of coefficient of variation. Otherwise, variability may assessed qualitatively (e.g., low, medium, or high), based on the evaluator's judgment.

[14] There are, of course, other considerations. See Section 5, Additional Considerations, for further discussion.

DRAFT

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 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.

DRAFT

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.[15] 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), and
- 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. Since residential populations tend to be relatively homogeneous, simple random sampling (SRS) is the most common sample design in this sector.

Commercial and industrial populations are composed of multiple subsectors (e.g., retail, office, grocery, manufacturing, and food processing). Nonresidential portfolios generally offer both prescriptive and custom measures for these sectors. Since 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 *ex ante* 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).

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

[15] This list is not exhaustive. Other possible segments include: low-income, agricultural, public/institutional, and transportation.

Appendix B). 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 5 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 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 shower heads and compact fluorescent lamps satisfy this requirement, but attic insulation does not, since 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 the following: (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

$$\boldsymbol{\frac{(1-p)}{p} \geq \frac{(1-0.4)}{0.4} = 1.5},$$

   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 these:

- Installation rates (e.g., proportion of program-provided CFLs installed);
- Measure Coverage (e.g., percent of insulation installed); and
- End-use parameters (e.g., 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).

When measure-level savings must be estimated with a prescribed level of precision and confidence, the sample size formula for the ratio estimator is this:

$$n = \left(\frac{z}{e_{\text{rel.}}}\right)^2 \left(\frac{s^{(\text{ratio})}}{\bar{y}}\right)^2$$

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 *ex ante* savings (see Appendix B).

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 *ex ante* 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 *ex ante* savings of the ith 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 *ex ante* 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. Since total savings is estimated as the realization rate multiplied by the *ex ante* 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.

Since 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 $\boldsymbol{N \cdot s^{(ratio)}/\sqrt{c}}$.

[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 amount 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 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:

$$\mathrm{CV} = \frac{s}{\bar{x}} = \frac{5.4}{18} = 0.3$$

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 kW draw. Meter-based estimates are then used to evaluate *ex post* 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.

DRAFT

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 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:

$$n_{\text{retail}} = 50.5 \cdot \left(\frac{25\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 12.6$$

$$n_{\text{office}} = 50.5 \cdot \left(\frac{21\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 10.6$$

$$n_{\text{rest.}} = 50.5 \cdot \left(\frac{9\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 4.5$$

$$n_{\text{school}} = 50.5 \cdot \left(\frac{12\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 6.1$$

DRAFT

$$n_{\text{light mfg.}} = 50.5 \cdot \left(\frac{33\%}{25\% + 21\% + 9\% + 12\% + 33\%}\right) \approx 16.7$$

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.[16] Then for sector *h*, the share of the sample will be proportional to:

$$\begin{aligned} N_h \cdot \frac{s_h}{\sqrt{c_h}} \quad &\propto \quad N_h \cdot \frac{[ex\ ante \text{ total for stratum } h]/N_h}{\sqrt{c_h}} \\ &= \quad \frac{[ex\ ante \text{ total for stratum } h]}{\sqrt{c_h}} \\ &\propto \quad \frac{[\text{stratum } h\text{'s percent of the } ex\ ante \text{ total}]}{\sqrt{c_h}} \end{aligned}$$

Also, evaluation costs differ among subsectors; engineers estimate the following hours are required to evaluate a site for each subsector:

Table 2. Evaluation Times and *Ex Ante* Savings by Subsector

Subsector	Hours	Proportion of *Ex Ante* Savings
Retail	2	25%
Office	4	21%

[16] To be precise, assume that within each stratum, the standard deviation of savings is proportional to the stratum's *ex ante* 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.)

Subsector	Hours	Proportion of *Ex Ante* Savings
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. *Ex Ante* Savings by Sector

Sector	*Ex Ante* kWh Total
Residential	2,900,000
Institutional	2,200,000
Commercial	3,300,000
Industrial	3,000,000
Total	11,400,000

DRAFT

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	*Ex Ante* 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 the Measure- and Site-Level Evaluation Planning Section.) Then use the cost-optimized allocation formula to determine sample the 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 *Ex Ante* kWh	*Ex Ante* 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 *ex ante* 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 based 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	*Ex Ante* kWh Total	*Ex Ante* Standard Deviation	Sample Size	Standard Error (*Ex Post* 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 *ex ante* 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	*Ex Ante* 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]

DRAFT

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.[17]

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.[18] 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 trade-off 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.

The general precision equation can be written in this form:

---

[17] 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.

[18] 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.

DRAFT

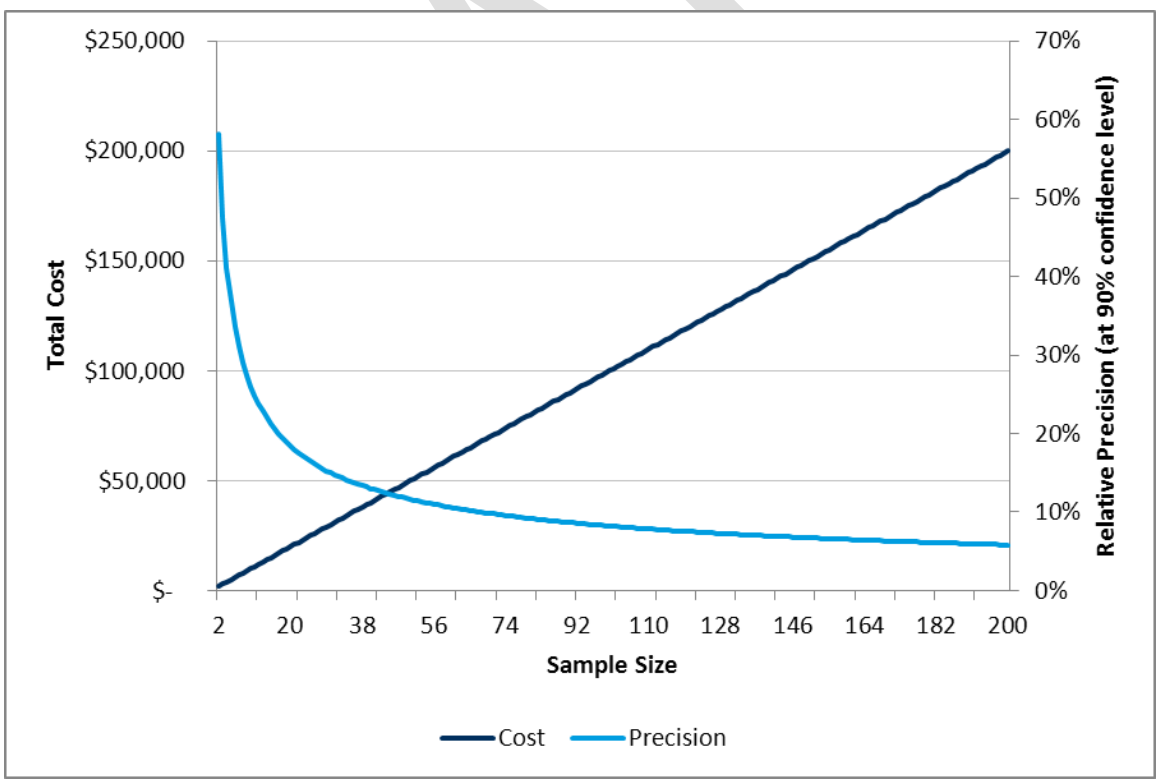
$$\text{Precision} = \text{confidence level}\sqrt{\frac{\text{variance}}{\text{sample size}}}$$

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.

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.

DRAFT

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 instance, 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** (i.e., not occurring by chance), such as non-coverage, non-response, self-selection, and some types of measurement errors; and
- **Random error** (i.e., 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.[19]

The distinction between systematic and random sources of error is important because different procedures are required to identify and mitigate each. While the amount of random error can typically be estimated using statistical tools, other means are required to estimate the level of systematic error. Since additional investment in the estimation process can lead to reductions in both types of error, trade-offs 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 (e.g., errors in recording data) may also cause this type of error. Metering results can be biased by equipment failure, incorrect placement, or poor calibration.[20] 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.

   Measurement error is reduced by investing in more accurate measurement technology, establishing clear data collection protocols, and by reviewing data to confirm they were accurately recorded. In most applications, this error source is

[19] Note that measurement error may be systematic or random. For example, a meter is not properly calibrated and consistently under or over estimates a measurement exhibits systematic error. A meter that is only accurate within a given interval is said to have random error within that interval.

[20] Such errors will bias measurements within a site. However, since 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.

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 instance, 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 (e.g., estimating the mean), and others are fairly complicated (e.g., 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

DRAFT

observed differences in energy use to temperature, while 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 which 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.[21]

3. **Regression.** Regression error may arise at either the measure/site level, or at the population/stratum level.

[21] ASHRAE Guideline 14-2002 and Guideline 2002R offer extensive guidance on accounting for measurement error. Also, see Section 1.8.5 of this document for a related discussion.

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 the last section 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 calculated with standard regression-related methods. Since 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 were steps taken to 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?

DRAFT

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.
- Also 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.
- Additionally, 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 lies 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.

DRAFT

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.[22]

Many research questions can be phrased in terms of either:

- A population *average*, such as average savings among program participants or proportion of participants with gas heat, or
- 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. Since 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*; and

n = the sample size.

To quantify the uncertainty surrounding this estimate, calculate the standard error and then calculate the precision.

[22] 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 under Simple Random Sampling in Appendix C.

The standard error of a proportion is most often[23] 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:

[23] 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 $\hat{p} = 92\%$. Since 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.[24]

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.

[24] 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

=`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,[25] the population average is estimated by the sample mean,

$$\bar{x} = \frac{\sum x_i}{n}$$

To quantify the uncertainty surrounding this estimate, calculate the standard error and then the precision. The sample mean's standard error is:

[25] Auxiliary information is discussed in the next section.

$$\widehat{\mathrm{SE}}(\bar{x}) = \frac{s}{\sqrt{n}}$$

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 *ex post* savings total, and the program database includes *ex ante* savings estimates for each member of the population.

For commercial and industrial projects, *ex ante* savings values often incorporate site-specific information, such as square footage of conditioned space and hours of operation. In these cases, *ex ante* values vary from project to project and the values can reasonably be expected to correlate with *ex post* savings values.

DRAFT

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[26]

$$\hat{\mu}_y = \frac{\sum y_i}{\sum x_i} \cdot \mu_x$$

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.

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.[27] This deficiency is especially pronounced with weighted estimators, since design-based weights describe selection probabilities (see Appendix C), 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:[28]

$$\text{Standard error of realization rate} = \widehat{\text{SE}}(\hat{r}) = \frac{1}{\sqrt{n}} \sqrt{\sum \frac{(y_i - \hat{r} \cdot x_i)^2}{\bar{x}^2 \cdot (n-1)}}$$

Thus, the standard error of the ratio-based estimate of μ_y is:

[26] 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.

[27] 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.

[28] 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.

$$\widehat{\mathrm{SE}}(\hat{\mu}_y) \;=\; \widehat{\mathrm{SE}}(\hat{r}\cdot\mu_x) \;=\; \widehat{\mathrm{SE}}(\hat{r})\cdot\mu_x \;=\; \frac{1}{\sqrt{n}}\cdot\frac{\mu_x}{\bar{x}}\cdot\sqrt{\sum\frac{(y_i-\hat{r}\cdot x_i)^2}{n-1}}$$

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 *ex post* 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 *ex post* total's relative precision expression has cancelling factors of μ_x in its numerator and denominator.

DRAFT

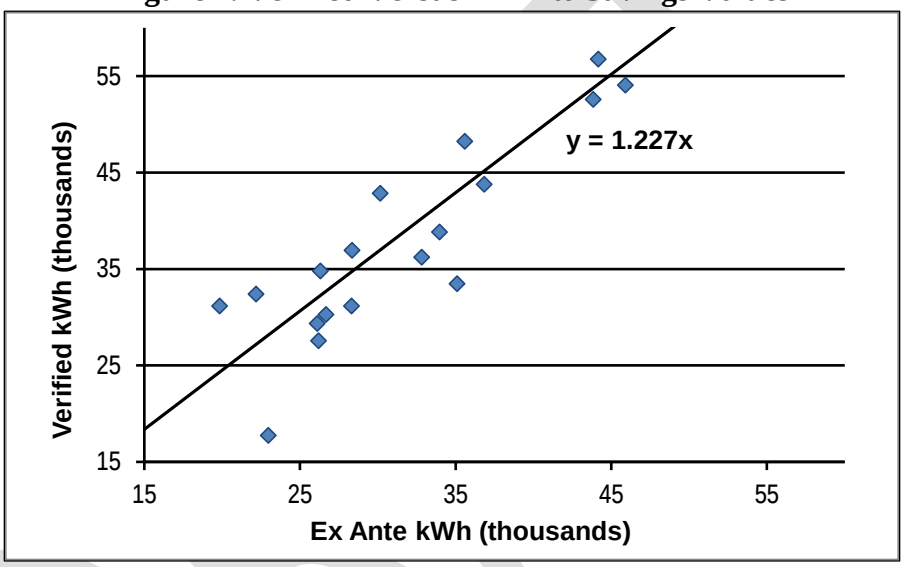
Example B-4

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 *ex ante* and *ex* post savings estimates.[29]

- The *ex ante* total for the sampled sites is 607,415 kWh, and
- The *ex post* 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 *Ex Ante* 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 (*ex ante*) 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:

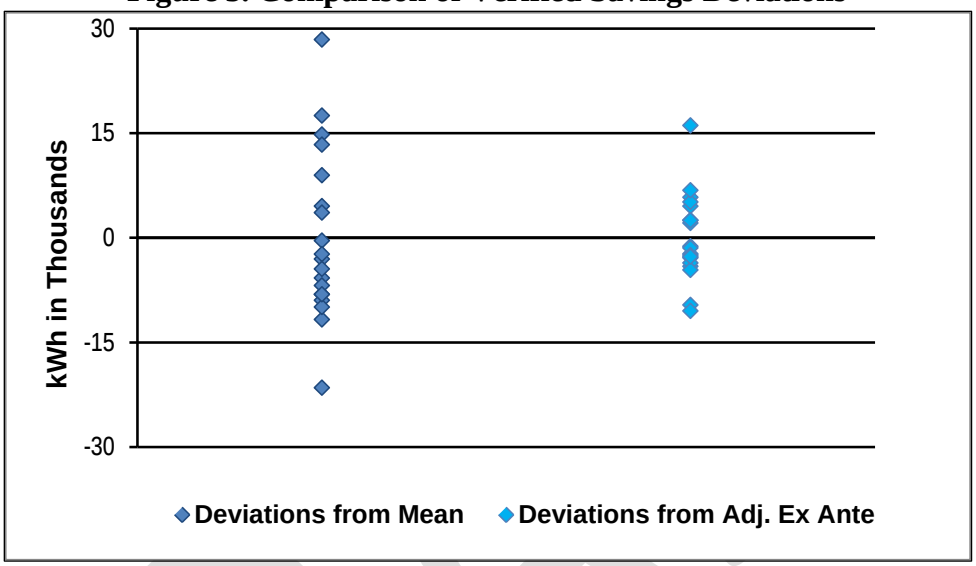
$$\text{Relative Precision}(N \cdot \bar{y}) \;=\; 1.645 \cdot \frac{12{,}132\ /\sqrt{20}}{39{,}216} \;=\; 11.4\%$$

[29] *Ex ante* values are the values in the program database, and *ex post* values are engineering estimates based on data collected on-site during the evaluation.

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 *ex ante* 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.

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 *ex ante* 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 *ex ante* savings.[30] (Also, most evaluators would agree that an ER of 15.7% and a CV of 30.9% are quite small for a commercial program.)

[30] 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).

Example B-5

The program database for a commercial gas efficiency program indicates 9.42 million Mcf of claimed (*ex ante*) savings program-wide, so we will conduct 40 sites 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:

$$\begin{aligned} \hat{r} &= 1.7\text{ mMcf}/2.0\text{ Mcf} &&= 85.0\% \\ \bar{y} &= 1.7\text{ mMcf}/40 &&= 0.0425\text{ mMcf} \\ \bar{x} &= 2.0\text{ mMcf}/40 &&= 0.0500\text{ mMcf} \end{aligned}$$

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:[31]

$$\text{Relative Precision}(\hat{r} \cdot \mu_x) = 1.645 \cdot \frac{s^{(\text{ratio})}/\sqrt{n}}{\bar{y}} = 1.645 \cdot \frac{0.0233/\sqrt{40}}{0.0425} = 14.3\%$$

So, we are 90% confident that the actual program savings is within 14.3% percent of 8.02 mMcf.

If we ignored the auxiliary (*ex ante*) 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 *ex post* and *ex ante* values are strongly correlated. For instance, if the CV in this example is 93.1%, then the mean-based estimator would be much less precise:

[31] Recall that the relative precision of the population *total* estimate is the same as the relative precision of the population *mean* estimate, since both of the estimates and their standard errors differ by a factor of *N* from one setting to the other.

$$\text{Relative Precision}(N \cdot \bar{y}) \;=\; 1.645 \cdot 0.931 \cdot \frac{1}{\sqrt{40}} \;=\; 24.2\%$$

[End of Example]

7.4 Estimating a Difference or Sum

Sums and differences of estimated quantities arise frequently in evaluation work. Two prominent examples are these:

- ***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:

$$\text{SE}(\hat{x} \pm \hat{y}) \;=\; \sqrt{\text{SE}(\hat{x})^2 + \text{SE}(\hat{y})^2 + 2 \cdot \text{Cov}(\hat{x}, \hat{y})}$$

Here, $\text{Cov}(\hat{x}, \hat{y})$ 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:

$$\text{SE}(\hat{x} \pm \hat{y}) \;=\; \sqrt{\text{SE}(\hat{x})^2 + \text{SE}(\hat{y})^2}$$

When the sampling errors are not independent, the evaluator will either need to estimate the covariance[32] or employ an alternate method, such as the bootstrap.

The absolute and relative precision are then estimated as:

$$\text{Absolute Precision}(\hat{x} \pm \hat{y}) \;=\; z \cdot \widehat{\text{SE}}(\hat{x} \pm \hat{y})$$

$$\text{Relative Precision}(\hat{x} \pm \hat{y}) \;=\; z \cdot \left(\frac{\widehat{\text{SE}}(\hat{x} \pm \hat{y})}{\hat{x} \pm \hat{y}} \right)$$

[32] The procedure for evaluating the covariance will depend on the particular estimators and their relationship to one another.

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 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[33], but when the two estimators' sampling errors are independent, the standard error is:

$$\mathrm{SE}(\hat{x} \cdot \hat{y}) = \sqrt{(\hat{x} \cdot \mathrm{SE}(\hat{y}))^2 + (\hat{y} \cdot \mathrm{SE}(\hat{x}))^2 + (\mathrm{SE}(\hat{x}) \cdot \mathrm{SE}(\hat{y}))^2}$$

[33] 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.

DRAFT

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 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.

Since 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)*	$\hat{x} \pm \hat{y}$	$\sqrt{\text{SE}(\hat{x})^2 + \text{SE}(\hat{y})^2}$	Either
Product*	$\hat{x} \cdot \hat{y}$	$\sqrt{(\hat{x} \cdot \text{SE}(\hat{y}))^2 + (\hat{y} \cdot \text{SE}(\hat{x}))^2 + (\text{SE}(\hat{x}) \cdot \text{SE}(\hat{y}))^2}$	Either

*The indicated standard error formula is only valid if estimators are statistically independent (see the previous two subsections).

DRAFT

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, simple random sampling without replacement (SRS) is the simplest random sampling approach, equivalent to "drawing *n* names from a hat."[34] 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 ith 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:

$$\bar{x} = \frac{1}{n} \sum_{\text{sampled } i} x_i$$

The standard error of the sample mean of an SRS is:

[34] 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.

$$\widehat{\mathrm{SE}}(\bar{x}) \;=\; \sqrt{1-\frac{n}{N}}\cdot\frac{1}{\sqrt{n}}\cdot\sqrt{\sum_{\text{sample}}\frac{(x_i-\bar{x})^2}{(n-1)}} \;=\; \sqrt{1-\frac{n}{N}}\cdot\frac{s}{\sqrt{n}}$$

[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.[35] 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

[35] 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 either 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:

$$n_0 = \left(\frac{z \cdot \mathrm{CV}}{e_{\text{rel.}}}\right)^2$$

Where:

CV is the coefficient of variation, the standard deviation divided by the mean

$e_{\text{rel.}}$ 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 since 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

DRAFT

normally-distributed, then a CV of 0.5 will apply.[36] This value, however, should not be applied 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.

Since 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:

$$\hat{p} = \frac{n_x}{n}$$

$$\hat{r} \cdot \mu_x = \frac{\sum y_i}{\sum x_i} \cdot \mu_x$$

Their standard errors, however, are multiplied by a finite population correction, just as in the sample mean case:

[36] 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 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 its members of the population will fall between 0 and 200% of the mean, then a CV of 0.5 is appropriate.

DRAFT

$$\widehat{\mathrm{SE}}(\hat{p}) = \sqrt{1-\frac{n}{N}} \cdot \frac{\sqrt{\hat{p}\cdot(1-\hat{p})}}{\sqrt{n}}$$

$$\widehat{\mathrm{SE}}(\hat{r}\cdot\mu_x) = \sqrt{1-\frac{n}{N}} \cdot \frac{1}{\sqrt{n}} \cdot \sqrt{\sum \frac{(y_i-\hat{r}\cdot x_i)^2}{n-1}} = \sqrt{1-\frac{n}{N}} \cdot \frac{s^{(\mathrm{ratio})}}{\sqrt{n}} \cdot \frac{\mu_x}{\bar{x}}$$

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:

$$e_{\mathrm{abs.}} = z \cdot \widehat{\mathrm{SE}}(\hat{p}) = z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n_0}}$$

So the initial sample size formula is:

$$n_0 = \left(\frac{z}{e_{\mathrm{abs.}}}\right)^2 \cdot p(1-p)$$

In this formula, z is as before and $e_{\mathrm{abs.}}$ is the absolute precision target. If there is no basis for making *a priori* assumptions about p, then use $p = 0.5$, since $p(1-p)$ 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.

DRAFT

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)**.[37] 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:

$$n = \frac{n_0 \cdot N}{n_0 + N}$$

[37] The California Evaluation Framework prescribes a model-assisted approach, based on evidence that deviations between *ex post* values y_i and adjusted *ex ante* 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.

8.1.3 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 *ex ante* 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 $N_1, N_2, \ldots, N_H$. Also assume that each stratum's sample is selected via simple random sampling within the stratum.[38] For example, within stratum h, an SRS of size n_h is been drawn from a group of N_h individuals, so each

[38] 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 which 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 AC tonnage among all commercial retailers in a particular service territory. A sample mean applied to a simple random sample would be very inefficient, since 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:

$$\hat{X}_{\mathrm{all}}^{(\mathrm{w})} = \hat{X}_1 + \hat{X}_2 + \hat{X}_3 = N_1 \cdot \bar{x}_1 + N_2 \cdot \bar{x}_2 + N_3 \cdot \bar{x}_3$$

The superscripted "w" emphasizes that this is a weighted estimator. Its standard error is:

$$\mathrm{SE}\left(\hat{X}_{\mathrm{all}}^{(\mathrm{w})}\right) = \sqrt{\mathrm{SE}(\hat{X}_1)^2 + \mathrm{SE}(\hat{X}_2)^2 + \mathrm{SE}(\hat{X}_3)^2}$$

To estimate the population-wide mean, use:

$$\hat{X}_{\mathrm{all}}^{(\mathrm{w})}/(N_1 + N_2 + N_3).$$

This estimate's standard error is:

$$\mathrm{SE}(\hat{X}_{\mathrm{all}}^{(\mathrm{w})})/(N_1 + N_2 + N_3).$$

[End of Example]

The general formula for the stratified-means estimator of the population total is:

$$\hat{X}_{\mathrm{all}}^{(\mathrm{w})} = \sum_{h=1}^{H} \hat{X}_h = \sum_{h=1}^{H} N_h \cdot \bar{x}_h$$

This estimator can also be written as a weighted sum,

$$\hat{X}_{\mathrm{all}}^{(\mathrm{w})} = \sum_{\text{sampled } i} \frac{N_{h(i)}}{n_{h(i)}} \cdot x_i = \sum_{\text{sampled } i} w_{h(i)} \cdot x_i$$

The weighted sum's standard error is calculated as follows. (Notice that only the within-stratum standard deviations, s_h, affect the standard error.)

DRAFT

$$\mathrm{SE}\left(\hat{X}_{\mathrm{all}}^{(\mathrm{w})}\right) = \sqrt{\sum \mathrm{SE}\left(\hat{X}_h\right)^2} = \sqrt{\sum N_h^2 \cdot \mathrm{SE}(\bar{x}_h)^2} = \sqrt{\sum \frac{N_h^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot s_h^2}$$

To estimate the population *mean*, divide the estimated total by the population size:

$$\bar{x}_{\mathrm{all}}^{(\mathrm{w})} = \frac{\hat{X}_{\mathrm{all}}^{(\mathrm{w})}}{N_{\mathrm{all}}}$$

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, $p_{\mathrm{all}} = N_{\mathrm{all}}^x/(N_1^x + N_2^x + \cdots + N_H^x)$

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, $\hat{p}_h = n_h^x/n_h$.

$\hat{p}_{\mathrm{all}}^{(\mathrm{w})}$ and $\widehat{N}_{\mathrm{all}}^x$ are our estimates of p_{all} and N_{all}^x.

The weighted estimators related to population proportions are:

$$\widehat{N}_{\mathrm{all}}^x = \sum_{h=1}^{H} N_h \cdot \hat{p}_h$$

$$\widehat{\mathrm{SE}}\left(\widehat{N}_{\mathrm{all}}^x\right) = \sqrt{\sum N_h^2 \cdot \widehat{\mathrm{SE}}(\hat{p}_h)^2} = \sqrt{\sum \frac{N_h^2}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \cdot \hat{p}_h(1 - \hat{p}_h)}$$

$$\hat{p}_{\mathrm{all}}^{(\mathrm{w})} = \frac{\widehat{N}_{\mathrm{all}}^x}{N_1 + N_2 + \cdots + N_H} = \frac{\sum N_h \cdot \hat{p}_h}{N_1 + N_2 + \cdots + N_H}$$

$$\widehat{\mathrm{SE}}\left(\hat{p}_{\mathrm{all}}^{(\mathrm{w})}\right) = \frac{\widehat{\mathrm{SE}}\left(\widehat{N}_{\mathrm{all}}^x\right)}{N_1 + N_2 + \cdots + N_H}$$

DRAFT

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:[39]

$$\begin{aligned}
\text{SE}\left(\hat{Y}_{\text{all}}^{(\text{w})}\right) &= \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right) \cdot \sqrt{\sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \sum_{\substack{\text{stratum } h \\ \text{sample}}} \frac{\left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2}{n_h - 1}} \\
&\approx \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right) \cdot \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{n_h}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \sum_{\substack{\text{stratum } h \\ \text{sample}}} \left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2} \\
&= \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right) \cdot \sqrt{\sum_{\text{sample}} w_{h(i)}\left(w_{h(i)} - 1\right)\left(y_i - \hat{r}_{\text{all}}^{(\text{w})} \cdot x_i\right)^2}
\end{aligned}$$

Typically, $\mu_{\text{all}}/\bar{x}_{\text{all}}^{(\text{w})}$ will be close to one, since 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

[39] See Särndal 1992, page 181.

now replaced by:

$$\left(s_h^{(r,\,\mathrm{w})}\right)^2 = \sum_{\substack{\text{stratum } h \\ \text{sample}}} \frac{\left(y_i - \hat{r}_{\mathrm{all}}^{(\mathrm{w})} \cdot x_i\right)^2}{n_h - 1}$$

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.

Table 10. Formulas for Stratified Estimators

Estimator	Expression	Standard Error
Weighted Sum	$\hat{X}_{\mathrm{all}}^{(w)} = \sum N_h \cdot \bar{x}_h = N \cdot \bar{x}_{\mathrm{all}}^{(\mathrm{w})}$	$\sqrt{\sum \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) s_h^2}$
Weighted Proportion	$\hat{p}_{\mathrm{all}}^{(w)} = \frac{\sum N_h \cdot \hat{p}_h}{\sum N_h}$	$\sqrt{\sum \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \hat{p}_h (1 - \hat{p}_h)}$
Weighted Ratio Estimator	$\hat{Y}_{\mathrm{all}}^{(r,\,\mathrm{w})} = \hat{r}_{\mathrm{all}}^{(r,\,\mathrm{w})} \cdot X_{\mathrm{all}}$	$\sqrt{\sum \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \left(s_h^{(r,\,\mathrm{w})}\right)^2} \left(\frac{\mu_{\mathrm{all}}}{\bar{x}_{\mathrm{all}}^{(\mathrm{w})}}\right)$

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	$\left(s_h^{(r,\,\mathrm{w})}\right)^2 = \sum_{\text{sample } h} \frac{\left(y_i - \hat{r}_{\mathrm{all}}^{(\mathrm{w})} \cdot x_i\right)^2}{n_h - 1}$	$\hat{r}_{\mathrm{all}}^{(\mathrm{w})} = \frac{\bar{y}_{\mathrm{all}}^{(\mathrm{w})}}{\bar{x}_{\mathrm{all}}^{(\mathrm{w})}}$

DRAFT

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.[40]

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:

[40] This is especially relevant for ratio estimators, since large deviations between *ex post* and *ex ante* values often reflect problems in the program database, rather than variation in consumer behavior.

$$\pi_i = \frac{n}{N}$$

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. Since 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 *ex ante* savings estimate from the program database.)
>
> y_i is the variable of primary interest for site *i*.

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.[41]

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*.[42] 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, since 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:

---

41 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).

42 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.

$$\hat{Y} = \sum_{\text{sampled } i} \frac{y_i}{\pi_i}$$

This estimator's standard error is estimated as:

$$\widehat{\text{SE}}(\hat{Y}) = \sqrt{\sum_{\text{sampled } i} (1 - \pi_i) \left(\frac{y_i}{\pi_i}\right)^2}$$

[End of Example]

Other PPS variants are available (see Särndal, *et al*, pp. 85-99).

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.[43]

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:

$$\hat{X}_h = N_h \cdot \bar{x}_h \quad \text{and} \quad \text{SE}(\hat{X}_h) = \sqrt{\frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) s_h^2}$$

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 $x_{i,1}, x_{i,2}, \ldots, x_{i,m}$ for some sample of measures within site *i*. The

[43] The distinguishing feature of two-stage sampling is that a sample of secondary units (e.g., measures) is selected within each sampled primary unit (e.g., 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 (e.g., phone surveys), and the initial sample data are used to increase efficiency for a small sample involving more expensive interactions (e.g., site visits). (Two-phase sampling is discussed in Section8.7.)

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:

$$\mathrm{SE}\left(\hat{X}_H\right) = \sqrt{\frac{N_H^2}{n_H}\left(1 - \frac{n_H}{N_H}\right) s_H^2 \;+ \sum_{\text{sample } H} \mathrm{SE}(\hat{x}_i)^2}$$

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 $\mathrm{SE}(\hat{y}_i)$.44 For some sites, this may include within-site ratio estimation or a PPS estimator. In such cases, the evaluator may review *ex ante* savings assumptions on site and adjust *ex ante* values to reflect actual hours of use and similar inputs, provided that the adjustments are: (1) applied to sampled and non-samples measures alike, and (2) based on information that is equally available for sampled and non-sampled measures.

For example, if the *ex ante* 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-level *ex ante* values may be adjusted accordingly. The main requirement is that such adjustments be made without giving the site's sampled measures any special consideration.[45]

Also, since *ex ante* values cannot be adjusted for every site in the population, this sort of *a priori* adjustment applies only measures within a sampled site and only to the calculation of $\hat{y}_i$ and $\mathrm{SE}(\hat{y}_i)$. The original *ex ante* 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 *ex ante* savings values from the program database (unadjusted) and the X_{all} is the *ex ante* total (unadjusted) for the entire population. The standard error is estimated as:

$$\widehat{\mathrm{SE}}\left(\hat{Y}_{\text{all}}^{(\text{w})}\right) = \left(\frac{\mu_{\text{all}}}{\bar{x}_{\text{all}}^{(\text{w})}}\right) \sqrt{\sum_{h=1}^{H} \frac{N_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)\left(s_h^{(r,\,\text{w})}\right)^2 + \left(\frac{N_H}{n_H}\right)^2 \sum_{\text{sample } H} \widehat{\mathrm{SE}}(\hat{y}_i)^2}$$

---

44 Recall that for ratio estimators, y_i represents verified savings and x_i represents *ex ante* savings estimates.

45 These *ex ante* adjustments need not be highly detailed, since 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 *ex ante* savings values.

Here, the standard errors of the $\hat{y}_i$ may reflect adjustments to measure-level *ex ante* 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 *ex ante* savings values for all sites in the Phase1 sample.
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 *ex ante* updates as auxiliary data.

In a two-phase study, the total savings is estimated as:

$$\hat{Y} = \hat{r} \cdot \hat{X} = \left(\frac{\sum_{\text{Sample 2}} y_i}{\sum_{\text{Sample 2}} x_i}\right) \cdot \left(N \cdot \frac{\sum_{\text{Sample 1}} x_i}{n_1}\right)$$

Since the *ex ante* values have been updated to reflect basic verification data, a large source of variation between *ex ante* and *ex post* 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:

$$\widehat{\text{SE}}(\hat{Y}) = N \cdot \sqrt{\left(1 - \frac{n_1}{N}\right)\frac{s_y^2}{n_1} + \left(1 - \frac{n_2}{N}\right)\frac{s_{\text{ratio}}^2}{n_2}}$$

Here, s_{ratio} calculated from the deviations between the updated *ex ante* values (Phase 1) and the final *ex post* 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.

DRAFT

EM&V Uniform Methods Project: Survey Design and Implementation Protocols for Estimating Gross Savings

Robert Baumgartner, Tetra Tech

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 the References and Resources section, as well as other sources that cover the specific topics in greater depth.

1 Introduction

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), include 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 freeridership, spillover, and market effects, they provide the basis for calculating net savings. (These applications are described in a separate chapter.)

In defining and describing best practices for survey research, the American Statistical Association states: "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."[1] 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; and
- 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.

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 total survey error within the constraints of a given research budget (Groves, 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).

[1] American Statistical Association, 1980.

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 (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 these: 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 a separate chapter.)

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 (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 SIC codes, since 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 the following: 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, 2011). The likelihood of a household being "wireless only" relates to a number of demographic characteristics, such as these:

- 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); and
- 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; and
- 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 the chapter titled, "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 that 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, 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 either 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 three factors correlating with errors in recall of retrospective events or behaviors:

- **Intervening events.** As related events occur—or new information related to the original event is obtained—individuals may 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.[2] 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, 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.

[2] Eisenhower, 1991

- Interviewers can administer surveys differently to different respondents. For example, interviewers may do the following: (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 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, 2009)

2.5.3.1.3 Respondents' Interpretation of Questions

Since 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 (such as, "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); and
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; and
- 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, 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, 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; and
- 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); or
- 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 utilize 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, 1999) Study findings indicate that applying these two techniques improves the results:

- Using words to anchor end-points and perhaps mid-points; and
- 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, 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, 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, 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, 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, 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; or

- 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, 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; and
- 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 these:

- The longer time required to complete data collection;
- The logistical difficulty of quality control measures, such as observing interviewers' conducting the interviews; and
- The potential for interviewer effects resulting from the 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 these:

- 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; and
- 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 these:

- 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; and
- 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 graphic, 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, 2009) It has long been a practice to mix modes in:

- The survey's contact phase (for or example, using an advance letter to contact respondents for telephone surveys or face-to-face interviews); and
- 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, 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.

The 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; and
- 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, this 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 the references section 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; and
- Minimum guidelines for reporting and disclosure of survey research.

6 References and Resources

American Association for Public Opinion Research. Website (www.aapor.org/Best_Practices1.htm).

American Statistical Association. 1980. *What Is a Survey?* Washington, DC.

Biemer, Paul, and Robert M.Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Seymour Sudman (eds.) 1991. *Measurement Errors in Surveys*. John Wiley & Sons.

Blumberg, Dr. Stephen J., and Julian V. Luke. 2001. "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.

Bradburn, Norman, Seymour Sudman, and Brian Wansink. 2004. *Asking Questions: The Definitive Guide to Questionnaire Design – For Market Research, Political Polls, and Social and Health Questionnaires.* John Wiley & Sons.

Christian, L. M., and D. A. Dillman, J. D. Smyth. 2008. "The Effects of Mode and Format on Answers to Scalar Questions in Telephone and Web Surveys." J. Lepkowski, C. Tucker, M. Brick, E. de Leeuw, L. Japec, P. Lavrakas, M. Link, and R. Sangster (eds.) Advances in Telephone Survey Methodology. Wiley-Interscience.

Converse, Jean M., and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire.* Sage Publications.

Couper, Mick P. 2011. "The Future of Modes of Data Collection." *Public Opinion Quarterly.* Vol. 75, No. 5, Pages 889-908.

de Leeuw, Edith D. 2005. "To Mix or Not to Mix Data Collection Modes in Surveys." Journal of Official Statistics, Vol. 21, No. 2, pages 233-255.

Dillman, Don A., and Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell, Jodi Berck, Benjamin Messer. 2009. "Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR), and the Internet." Social Science Research, Vol. 38, pages 1-18.

Eisenhower, Donna, and Nancy A. Mathiowetz, David Moganstein. 1991. "Recall Error: Sources and Bias Reduction Techniques." Biemer, Groves, Lyberg, Mathiowetz, and Sudman (eds.) *Measurement Errors in Surveys*. John Wiley & Sons.

Groves, Robert M., and Lars Lyberg. 2010. "Total Survey Error: Past, Present, and Future." Public Opinion Quarterly, Vol. 74, No. 5, pages 849-879.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. Wiley Series in Survey Methodology. Wiley-Interscience: New York.

Holbrook, A. L., and J. A. Krosnick. 2010. "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item-Count Technique." Public Opinion Quarterly, Vol. 74, No. 2, pp. 328–343.

Krosnick, Jon A. 1991. "Response Strategies for Coping with Cognitive Demands of Attitude Measurement in Surveys." *Applied Cognitive Psychology*, Vol. 5, pages 213-236.

Krosnick, Jon A., and Stanley Presser. 2009. "Question and Questionnaire Design." James D. Wright and Peter V. Marsden (eds.) *Handbook of Survey Research (2nd Edition).* Elsevier: San Diego.

Krosnick, Jon A., and John P. Robinson, Philip R. Shaver, and Lawrence Wrightsman (eds.). 1999. "Maximizing Questionnaire Quality." *Measures of Political Attitudes*. Academic Press, San Diego.

Madans, Jennifer, Kristen Miller, Aaron Maitland, and Gordon Willis. 2011. *Question Evaluation Methods: Contributing to the Science of Data Quality.* John Wiley and Sons.

Nunnally, Jum C. 1978. *Psychometric Theory* (2nd ed.) McGraw-Hill, New York.

O'Muircheartaigh, Colm, and Jon Krosnick, A. Helic. 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.

Schuman, Howard, and Stanley Presser. 1996. *Questions and Answers in Attitude Surveys*: *Experiments on Question Form, Wording, and Context*. Sage Publications.

Schuman, Howard, and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context.* Academic Press.

Schwartz, Norbert. 1999. "Self-Reports: How the Questions Shape the Answers." American Psychologist, Vol. 54, No. 2, pages 93-105.

Sudman, Seymour, and Norman Bradburn, Norbert Schwartz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. Jossey-Bass, San Francisco.

Visser, Penny S., and Jon A. Krosnick, Paul J. Lavrakas. 2000. "Survey Research." Harry T. Reis and Charles M. Judd (eds.) *Handbook of Research Methods in Social and Personality Psychology.* Cambridge University Press.

Wickman, A., and B. Warneryd. "Measurement Errors in Survey Questions." Social Indicators Research, Vol. 22, pages 199-212.

ASSESSING PERSISTENCE AND OTHER EVALUATION ISSUES

Daniel M. Violette, Navigant Consulting

Addressing other evaluation issues that have been raised in the context of energy-efficiency (EE) 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 EE programs and portfolios. In addition to discussing "persistence" (which refers to the stream of benefits over time from an EE measure or program), this chapter provides a summary treatment of these issues:

- Synergies across programs,
- Rebound,
- Dual baselines, and
- 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 Persistence of Energy Savings

Understanding persistence is critical to making good decisions regarding EE investments, so this section outlines program evaluation methods that can be employed to assess persistence—the reliability of savings over time. EE program benefits are measured as the net present value (NPV) of a stream of benefits based on the energy and demand savings[1] 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 EE 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

[1] 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.

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).

1.1 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.[2] 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 & Resource Solutions 2005).[3] In that study, overall persistence is divided into two components: (1) measure life, and (2) savings persistence.

Recognizing that definitions for the terms *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 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.

1.1.1 Definitions

The definitions of key terms used in this chapter are these.

1.1.1.1 Measure Life or Effective Useful Life (EUL)

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.

---

[2] 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.

[3] 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.

- “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.

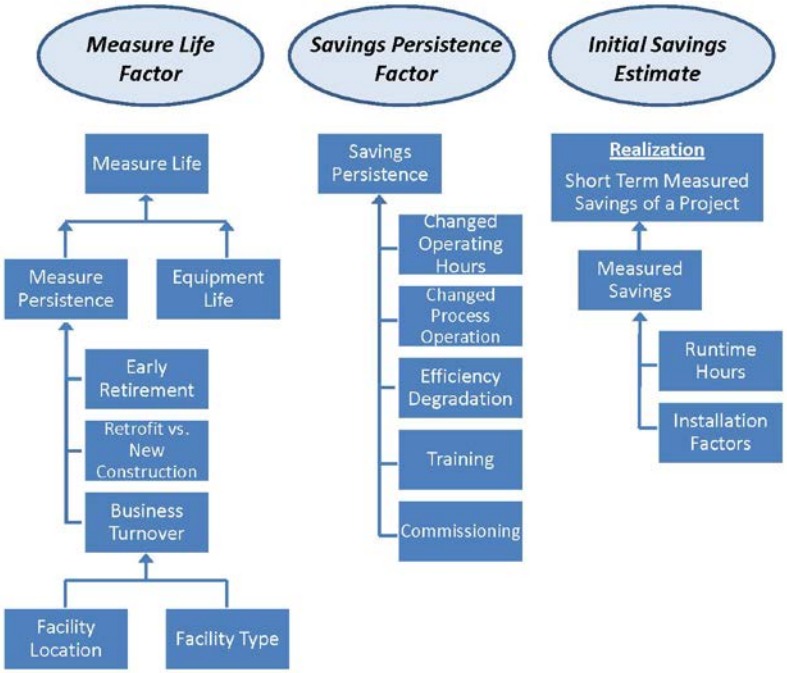
1.1.1.2 Savings Persistence

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).

DRAFT

Figure 1. Relationship of Measure Life, Savings Persistence, and Initial Savings Estimates[4]



1.1.2 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

1.1.2.1 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.[5] 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.

---

4 Source: *Adapted from Energy &Resource Solutions (2005).*

5 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.

1.1.2.2 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.

1.1.2.3 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 7 and year 8. The impact of discounting depends on the discount rate being used and the measure life.[6]

1.1.2.4 Differences in Baseline and EE Energy Streams of Benefits.

Energy savings calculations are based on the difference between the post EE 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 *and* it would be appropriate to assume constant energy savings over the life of the EE measure.[7] In fact, if the relative persistence of savings is higher for the EE measures compared to a baseline consisting of standard measures, then energy savings not only persists, but can increase over time.

Those four factors are meant to address the following questions:

---

[6] 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.

[7] The report from Peterson et al. (1999) is a good example of degradation being measured for both an efficient appliance offered by an EE 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.

- 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 these common factors are listed in Table 1.

Table 2. Factors Influencing Persistence

Residential Sector Programs and Measures	Commercial and Industrial Sector Programs and Measures
1. Changes in ownership 2. Maintenance practices 3. Changes in equipment use 4. Behavioral changes 5. Occupancy changes 6. Inappropriate installation of equipment 7. Manufacturer performance estimates that do not reflect in-field operating conditions.	1. Business turnover 2. Remodeling 3. Varying maintenance 4. Operating hours and conditions 5. Inappropriate installation of equipment 6. 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.

1.2 State of the Practice in Assessing Persistence

Professional judgment plays a significant role in selecting a method for assessing persistence. The California EE Evaluation Protocols (California Public Utilities Commission, 2006) has several types of retention, degradation, and measure 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[8] a database of information on measure life and performance degradation.

[8] 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

- ***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 is a large number of EE 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.

1.3 Database/Benchmarking Approaches

The three examples of database/benchmarking approaches presented below are based on:

- Engineering judgment
- Experience with the EE measures, and
- Information on local and regional conditions to develop tables of measure lives for use in EE 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.)

1.3.1 Example Study 1: 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 EE programs in the New England states that have supported this research effort;
- Satisfy any ISO-NE requirements (e.g., for definition and documentation sources); and,
- Work as common values, accepted by all New England states for the FCM (i.e., the ISO-NE forward capacity market).

Methodology: "Reviewed all secondary data collected and developed a preliminary list of potentially applicable residential and C&I measures. This list was then distributed to program administrator staff within the SPWG 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."

---

databases are sometimes the source of deemed values for measure life and persistence of savings that are used in evaluation efforts.

1.3.2 Example Study 2: 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.)

1.3.3 Example Study 3: Energy & Resource Solutions (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 EE programs
- Develop a table of technological measure lives
- Recommend common measure lives and persistence assumptions to be used by the sponsors."

Methodology: "ERS 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 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."

1.4 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 field work. However, new technologies and measures installed less frequently pose greater challenges for this judgment-based benchmarking approach. For many widely implemented EE 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 (2009, 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 comprise a program.

1.5 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 EE program generally rely on the following:

- 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.[9]
- Statistical analyses using regression-based methods to estimate retention/survival models that produce estimates of the survival or failure rates of EE measures.

The California EE Evaluation Protocols[10] 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.
- ***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.

---

[9] 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 chapter on Survey Research.

[10] The methodology language from the California EE 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.

- ***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 EE 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.

1.5.1 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 4 to be consistent with the retention study. This ratio for year 4 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.)

1.5.1.1 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

DRAFT

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.[11]

1.5.1.2 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 (and discussed in the Sample Design chapter).

1.5.1.3 Examples of Retention and Measure Life Studies

Two examples of these types of studies were performed by KEMA and by Nexus Market Research.

- KEMA (2004) used a telephone survey to gather information on refrigerators at years 4 and 9, as part of a review of an appliance recycling program.
- 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.[12] Each is discussed below.

Example Study 1: 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, and
- 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:

---

[11] In addition to this language, the California EE Evaluation Protocols outline certain sampling criteria that must be met in California. However, these criteria may vary in accordance with the requirements of different jurisdictions.

[12] To assist evaluators, the California EE Evaluation Protocols state: "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 under-estimating the measure life/EUL values *and* that the realization rate exceeds 1.0.

Figure 2. KEMA (2004) Table E-1

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

Example Study 2: Nexus Market Research (2008). This study examined the measure life of lighting products distributed through EE programs in New England.

The definition of measure life is the same as presented above in the Addressing Persistence section and used in Energy & 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 (i.e., 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 (i.e., measure persistence).

DRAFT

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 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)[13] and two for exterior fixtures (markdown and all other programs).

[13] 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).

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).

1.5.2 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 in1999.)
- Performing a 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.)

Example Study 1: 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. 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.

The results of 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. EE measures may have performance degradation, but so does standard equipment. If the EE 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 EE 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.[14]

The primary purpose of this research was to establish the technical degradation factors (TDF), 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.

Figure 3. Proctor Engineering (1999) Table ES-1

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

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.

[14] 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.

DRAFT

Example Study 2: 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 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 over 20,000 treatment customers and a control group of over 30,000 customers. Thus, large samples are needed to identify small effect sizes from EE programs.

This regression framework can be applied to a third and fourth year of data to assess longer-term participation.

1.6 Persistence Recommendations and Conclusions

Evaluators address the issue of persistence of savings from EE 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 used which comprise 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. Appendix A to this chapter presents tables outlining program and measure persistence study challenges and issues.

This 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 EE programs, and this requires both an understanding of the techniques and applied judgment.

1.6.1 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 EE 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 EE 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 EE programs. However, past work can be improved upon through the use of more systemized approaches, such as a Delphi-type of analysis.[15]
- 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.[16]

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.[17]

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.)

---

[15] 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 (2003).

[16] Billing data analyses that try to estimate small effects reliably (e.g., 2% savings) without the required sample sizes and accurate data for the independent variables (i.e., little measurement error) have often not been successful. Quantum (1998) discusses this issue in the context of using a billing analysis to assess persistence for new home construction.

[17] Panel data methods are suggested as a potential approach in both Skumatz (2012) and Nexus (2008).

2 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.

2.1 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 do the following:

1. Identify what they believe may be positive and negative synergies (i.e., direction); and
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 EE and future program/portfolio designs.[18]

2.1.1 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 EE program design considerations.[19]

---

[18] 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, e.g., 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 (2003), and a version of this method is discussed in EPRI (2010, p. 5-4). This information can be used by the program administrator to inform the design of future EE portfolios.

[19] 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

2.2 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 EE 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.

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. (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."

---

models that use different scenarios regarding the magnitude and direction of synergies can help assess the robustness of program and portfolio cost-effectiveness.

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.[20] 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 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 EE activities and changes in energy use.[21] 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.

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 (2011) and by 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

---

[20] The assumption is that observations of the independent variable can be considered fixed in repeated samples (i.e., that it is possible to repeat the sample with the same independent variable values; Kennedy, 2003, p.49).

[21] Two recent publications with examples of this are Rivers and Jaccard (2011) and Arimura et al. (2011).

set has many observations with missing values for DSM spending, even after our meticulous efforts to find them from various sources."[22]

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 EE 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.[23]

Attenuation bias can be an important problem in regression models using independent variables that might have large amounts of measurement errors due to:

- Differences in reporting of values in databases compiled across utilities, or
- Assignment/allocation of values at a utility service territory level down to a county level to create more observations.

Chen et al. (2007, 2011) and 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).

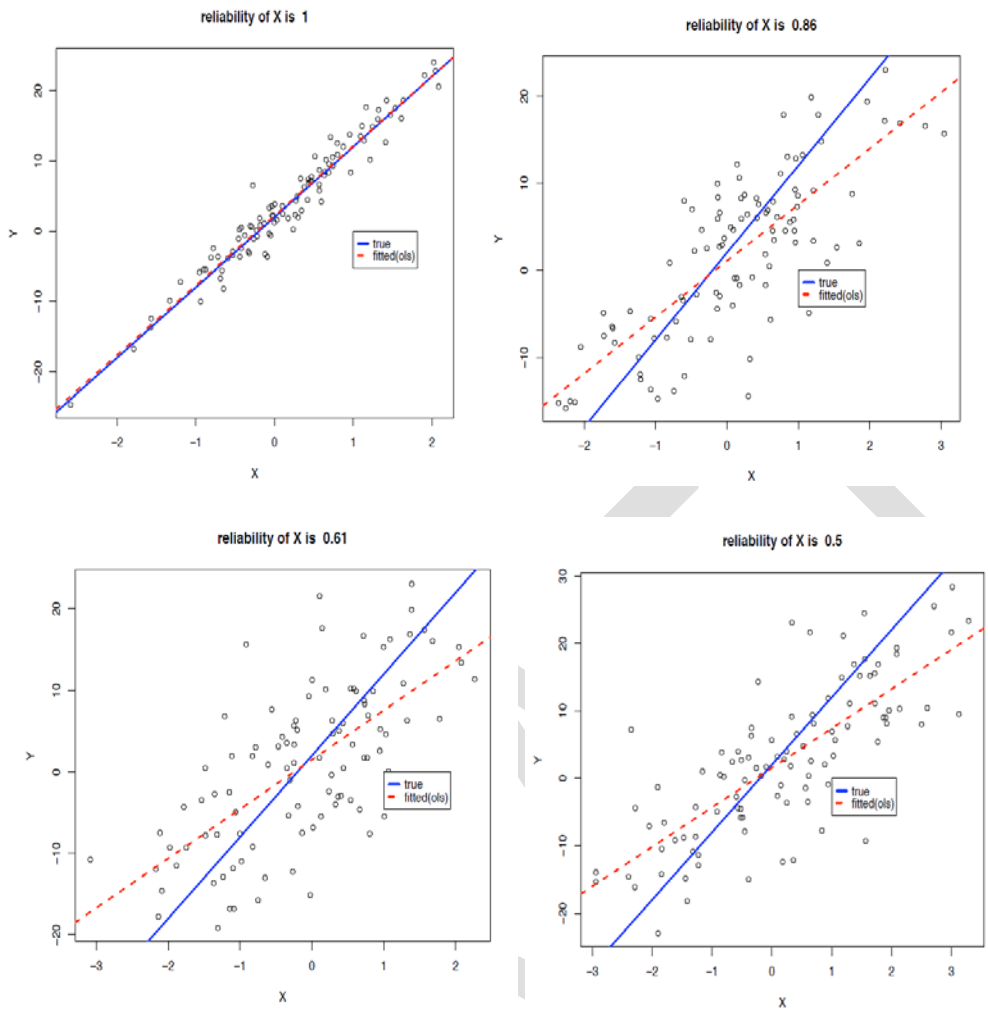
Each simulation is shown in Figure 5.

---

[22] 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 (i.e., the utility's EE expenditures that could explain changes in energy use and be used to assess cost-effectiveness in terms of cost per kWh saved).

[23] 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 (1957) shows that, because of the attenuation bias, the estimated influence of income on consumption would be underestimated.

Figure 1. Satorra (2008) Simulation Results



As shown in Figure 5, 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.

2.2.1 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 EE evaluation.

Nevertheless, having a good data platform from which EE savings are evaluated is important and needs more emphasis in practical evaluation work.

2.3 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 EE 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 EE measure; or
- The savings could be based on the difference between the new standard measures available in the market as compared to the new EE 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.

2.3.1 Conclusion

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 complexity and uncertainty entailed in estimate 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 EE 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.

2.4 Rebound Effects

Rebound occurs when the costs of using energy are reduced due to EE programs. When families spend less money to cool their home in the summer because of more efficient equipment, they might change their temperature setpoint 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 EE program measure—such as the installation of a high-efficiency air conditioner) then "take back" some of those savings by using more energy.[24]
- ***Type 2: Rebound takes place in the larger economy.*** EE 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 EE program evaluations.[25] 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-making.[26] (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 efficiency of the unit. In other words, consumers generally do not change their thermostat setting as a result of participating in an EE program.

---

[24] 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.

[25] This chapter is focused on EE 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 EE programs that influence all hours of a season.

[26] The primary reference for this concept is Simon (1957), but it is also discussed in Kahneman (2003).

Low-income customers can be the exception, as they may change their thermostat setpoints for both AC and heating after participating in an EE 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 EE 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 EE 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 EE 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.[27]

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.

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.

[27] The Rocky Mountain Institute has posted information at http://blog.rmi.org/blog_Jevons_Paradox discussing recent hypotheses about this type-two rebound effect. Other references are Tierney J. (2011), which presents the issue of rebound as being important, and a counterpoint paper by Afsah (2011).

3 References and Resources

Ahmad, M. , Deng, A., Spoor, S., Usabiaga, M., and I. Zhao (2011). *Persistence of Energy Savings in Industrial Retrocommissioning Projects*. ACEEE Summer Study on Energy Efficiency in Industry.

Allison, P.A. (1995).*Survival Analysis Using the SAS® System: A Practical Guide.* SAS® Institute.

Arimura, T.H., S. Li, R.G. Newell, and K. Palmer (2011). *Cost-Effectiveness of Electricity Energy Efficiency Programs.* National Bureau of Economic Research, NBER Working Paper No. 17556.

California Public Utilities Commission (2006). *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*. Oregon: TecMarket Works. Available online: http://www.calmac.org/publications/EvaluatorsProtocols_Final_AdoptedviaRuling_06-19-2006.pdf.

Chen, X., H. Hong, and D. Nekipelov (2007). Measurement Error Models. Prepared for the *Journal of Economic Literature*. Available online: http://www.stanford.edu/~doubleh/eco273B/survey-jan27chenhandenis-07.pdf.

Chen, X., H. Hong, and D. Nekipelov (2011). Nonlinear Models of Measurement Errors. *Journal of Economic Literature*, 49(4): 901–37.

Decision Sciences Research Associates (1999). *1994 Commercial CFL Manufacturers' Rebate Persistence Study.* Report ID 529D, Pasadena: Decision Sciences Research Associates.

Energy and Resource Solutions (2005). *Measure Life Study.* Prepared for The Massachusetts Joint Utilities, November.

EPRI (2010), *Methodological approach for Estimating the Benefits and Costs of Smart Grid Demonstration Projects*. EPRI, Palo Alto, CA: 2010. 1020342.

Friedman, M. (1957). *A Theory of the Consumption Function*. Princeton University Press.

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.

Gavankar, S. and R. Geyer (2010). The Rebound Effect: State of the Debate and Implications for Energy Efficiency Research. Institute of Energy Efficiency (UCSB), June. Available online: 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): 1449–75.

KEMA (2004). *Residential Refrigerator Recycling Ninth Year Retention Study*. Study ID Nos. 546B, 563, Madison. 2004.

KEMA (2006). *2005 Smart Thermostat Program Impact Evaluation*. Final Report prepared for San Diego Gas and Electric Company.

DRAFT

KEMA (2009). *Focus on Energy Evaluation: Business Programs: Measure Life Study Final Report. P*repared for the Public Service Commission of Wisconsin, August 25.

Kennedy, P. (2003). *A Guide to Econometrics.* Chapter 9, MIT Press.

Navigant Consulting, Inc. (2010). *Kaizen Blitz Pilot. Report 1.* Prepared for Energy Trust of Oregon.

Navigant Consulting, Inc. (2011). *Evaluation Report: OPower SMUD Pilot Year 2*. Presented to OPower, February.

Nexus Market Research, Inc. (2008). *Residential Lighting Measure Life Study*. Cambridge: New England Residential Lighting Program.

Ontario Power Authority (2011). *EM&V Protocols and Requirement*. Ontario: 2011-2014.

Pacific Gas & Electric (1999). *Commercial Lighting Study.* CALMAC, pp. 4-14.

Peterson, G., R. deKieffer, J. Proctor, and T. Downey (1999). *Persistence #3A: An Assessment of Technical Degradation Factors: Commercial Air Conditioners and Energy Management Systems, Final Report*. CADMAC Report # 2028P.

Rocky Mountain Institute (2011). *Jevons Paradox: The Debate That Just Won't Die*. By Potts, M. and C. Burns.

Proctor Engineering (1998). *Statewide Measure Performance study #2: An Assessment of Relative Technical Degradation Rates. Final Report.* For California Measurement Advisory Committee.

Proctor Engineering (1999). *Summary Report of Persistence Studies: Assessments of Technical Degradation Factors, Final Report*. CADMAC Report #2030P.

Quantum Consulting (1998). *PG&E Statewide Multi-Year Billing Analysis Study: Commercial Lighting Technologies Final Report*. Berkeley.

Rivers, N. and M. Jaccard (2011). Electric Utility Demand Side Management in Canada. *The Energy Journal*, 32 (4), pp. 93-116.

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 (1999). *Commercial Energy Efficiency Incentives Ninth Year Retention Evaluation.* 1996 and 1997.

Satorra, A. (2008). *Theory and Practice of Structural Equation Modeling, Department d'Economiai Empresa*. Barcelona. Available online: 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 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.*

DRAFT

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.

U.S. Environmental Protection Agency (2007). *Model Energy Efficiency Program Impact Evaluation Guide.*

Violette, D. and K. Cooney (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. December. Web: http://www.theboc.info/pdf/Eval-BOC_SummittBlue_NEEA_2003.pdf

DRAFT

3 Appendix A: (sent as a separate document – with edits couldn't get it in right)

DRAFT