NAVIGANT

A Sampling Methodology for Custom C&I Programs

Prepared for:
Sub-Committee of the
Technical Evaluation Committee

November 12, 2012
Revised: October 28, 2014

Prepared by:
Dan Violette, Ph.D. & Brad Rogers, M.S., MBA

Navigant Consulting, Inc.
1375 Walnut Street, Suite 200
Boulder, CO 80302
303.728.2500
www.navigant.com

Acknowledgements

The authors wish to acknowledge Leslie Kulperger and Meredith Lamb of Union Gas Limited and Judith Ramsay and Rod Idenouye of Enbridge Gas Distribution for their guidance, assistance, and support of this work. The authors also wish to acknowledge Chris Neme and Bob Wirtshafter for their thoughtful criticism and direction on earlier drafts of this report. The authors appreciate the opportunity to work with such a knowledgeable and discerning team.

NAVIGANT

Table of Contents

1. Introduction

This report presents a sampling methodology intended for use in the evaluation of custom demand side management (DSM) programs delivered in commercial and industrial (C&I) sectors. The report provides a technical explanation of issues that have been raised in the evaluation processes. It also provides justification for the approaches recommended herein.

Past evaluation studies of Union Gas Limited (Union) and Enbridge Gas Distribution (Enbridge) custom programs have undergone third-party audits where the sample design and realization rate calculations are examined. The processes and judgments applied in these evaluation studies are audited to ensure that the analyses are transparent and accurate. The recommendations in this report along with the technical discussions are intended to better frame the issues for the third-party audit reviews and streamline the overall audit process.

The sample design methodology recommendations are presented in Section 5. The realization rate and achieved precision methodology recommendations are presented in Section 6. The report also contains three technical appendices discussing key issues and presenting the calculations required to develop statistical program estimates.

1.1 Background

Union and Enbridge have delivered DSM initiatives since 1997 and 1995, respectively. Union and Enbridge operate DSM programs, including programs that involve custom projects in the industrial, commercial, multi-residential, and new construction sectors. Custom projects cover opportunities where savings are linked to unique building and manufacturing specifications, end uses, and technologies. Each project is assessed individually for participation in the program. The DSM portfolio for both utilities includes several hundred custom projects annually.

Union and Enbridge DSM activities are regulated by the Ontario Energy Board (OEB) and adhere to the requirements as laid out in DSM Guidelines for Natural Gas Utilities.[1] For custom projects, the resource savings are determined through engineering calculations that are determined at the design stage of each project. There is a need to verify the resource savings through a third-party C&I engineering review.

A sampling methodology for custom projects was developed in 2008.[2,3] This methodology was intended to be used to evaluate future custom program impacts while the programs retained

[1] "Demand Side Management Guidelines for Natural Gas Utilities." EB-2008-0346. Ontario Energy Board. June 30, 2011.

[2] "Sampling Methodology for Engineering Review of Custom Projects." Enbridge Gas Distribution Inc. and Union Gas Limited. Prepared by Summit Blue Consulting. April 3, 2008.

roughly the same distribution of projects in terms of size and segment. There have been some changes to the custom programs and Union and Enbridge are now preparing for the engineering review of custom projects for 2012. As a result, there is a need to update the sampling methodology. Both utilities seek a harmonized approach to evaluating custom programs that involves on-site reviews of selected custom projects within a representative sample of the respective utility project populations.

In 2012, both utilities entered into a new regulatory framework in Ontario that established a new intervener process with the creation of a common Technical Evaluation Committee (TEC) for both utilities. The goal of the TEC is to establish DSM technical and evaluation standards for natural gas utilities in Ontario. The TEC will make recommendations to the OEB on annual Technical Reference Manual (TRM) updates, establish evaluation priorities, and reach consensus on the design and implementation of evaluation studies.

1.2 OEB Requirements for Evaluating Custom Projects

The OEB's DSM Guidelines for Natural Gas Utilities draws special attention to custom projects. The Guidelines define custom projects:[4]

> *Custom projects are those projects that involve customized design and engineering, and where a natural gas utility facilitates the implementation of specialized equipment or technology not identified in the Board approved list of input assumptions. Projects that simply include a combination of several measures provided in the list of input assumptions are not considered to be custom projects.* (p.5)

The Guidelines go on to prescribe an evaluation approach for custom projects:

> *For custom resource acquisition projects, which usually involve specialized equipment, savings estimates should be assessed on a case by case basis. It is expected that each custom project will incorporate a professional engineering assessment of the savings. This assessment would serve as the primary documentation for the savings claimed.*
>
> *A special assessment program should be implemented for custom projects. The assessment should be conducted on a random sample consisting of 10% of the large custom projects; and the projects should represent at least 10% of the total volume savings of all custom projects. The minimum number of projects to be assessed should be 5. Where less than 5 custom projects have been undertaken, all projects should be assessed. The assessment should focus on verifying the equipment installation, estimated savings and equipment costs.*

[3] "Update Memorandum: Proposed Sampling Method for Custom Projects." Summit Blue Consulting. October 31, 2008.

[4] "Demand Side Management Guidelines for Natural Gas Utilities." EB-2008-0346.Ontario Energy Board. June 30, 2011.

> *All program result evaluations should be conducted by the natural gas utilities' third-party evaluator(s). If possible, the natural gas utilities' third-party evaluator(s) should be selected from the [Ontario Power Authority's] OPA's third-party vendor of record list. The natural gas utilities' third-party evaluators should seek to follow the OPA's evaluation, measurement and verification protocols,[5] where applicable and relevant to the natural gas sector.* (p.39)

The recommended sample methodology contained in Sections 5 and 6 of this report conforms to the Guidelines for custom projects. Appendix B presents the detailed equations necessary to implement the recommended methodology.

1.3 Report Objective

The objective of this report is to develop a methodology for designing a sample and for calculating achieved realization rates and sample confidence and precision using the observed results from the sample. The recommended methodology must meet OEB requirements as well as address the technical and programmatic needs of Union and Enbridge custom programs. The steps taken to achieve this objective include the following:

- Understand the composition of Union and Enbridge custom programs (Sections 2 and 3)
- Review and analyze sample methodologies in selected jurisdictions (Section 4)
- Recommend a methodology for designing and selecting samples (Section 5)
- Recommend a methodology for calculating the achieved program realization rates and sample confidence and precision (Section 6)

The recommended statistical methodology can be described as two-stage stratified ratio estimation. A step-by-step approach to implementing the methodology for sample design is presented in Section 5.4.

The recommended sample methodology is intended to provide sufficient flexibility to allow Union and Enbridge to efficiently meet sample precision needs while the composition, participation, and impacts of their custom programs resemble the current 2011/2012 programs. If the nature of the custom programs changes, adjustments to the recommended methodology may be warranted.

[5] "EM&V Protocols and Requirements: 2011-2014." Ontario Power Authority. March 2011. (see page 129)

2. Overview of Union Custom Programs

Union's T1/R100 and commercial/industrial (C/I) custom programs are aligned under one brand platform, the *EnerSmart* program. This ensures a seamless, recognizable brand throughout Union's franchise. The program scorecards are divided based on rate class.[6] The T1/R100 program consists of T1 rate customers in Union's Southern delivery zone whose annual consumption is over 5M m^3 and R100 rate customers in Union's other delivery zones whose annual consumption is over 25.6M m^3. The C/I program consists of Union customers in all other rate classes. The methodology in this report pertains only to the custom measures in these programs. Additionally, Union is adding a new Low Income custom segment for the 2012 program year.[7]

Figure 1 outlines the rate class divisions of Union's custom projects. The number of projects in the C/I program is more than twice the number of the projects in the T1/R100 program but represents less than half of the savings of that program.

Figure 1. Union 2011 Custom Projects Overview

Union Custom Sector	# of Custom Projects	Gas Savings	% of Custom Portfolio
T1/R100	200	98,702,955	68.3%
Commercial/Industrial	459	45,472,108	31.5%
Low Income*	13	348,525	0.2%
Total	**672**	**144,523,588**	**100%**

*Low Income values are forecast for 2012 as this is a new segment for Union in 2012.
Source: Union Gas Limited

Custom projects are highly heterogeneous, with most projects tied directly to unique processes or technology requirements. Each project is validated on a stand-alone basis by a comprehensive professional engineering review and the overall programs are required to pass a Total Resource Cost (TRC) screening process. The EnerSmart program was designed to achieve savings in process-specific energy applications, as well as space heating, water heating, and the building envelope. Given the customized nature by which tracking database savings estimates are generated, Union conducts a third-party, on-site engineering study to verify the results of a representative project sample.

Account managers market the program directly to customers for T1/R100 and a combination of directly and indirectly through trade allies, channel partners, energy service companies, engineering firms, and equipment manufacturers to all other rate classes. Account managers work to cost-effectively promote energy efficiency within Union's C&I customer base.

[6] Historically, the Union custom C&I program was divided based on whether the customer purchased gas under a firm distribution contract or through a general service contract.
[7] Low income includes commercial and industrial general service customers.

3. Overview of Enbridge Custom Programs

Enbridge offers custom programs for the C&I sectors. A variety of incentive-based initiatives are offered to C&I sector customers. These initiatives include custom project incentives and a suite of prescriptive offerings aimed at promoting specific measures. Given the myriad of building types, end uses, ownership structures, and leasing arrangements, the C&I sector is a complex and variable segment in which to market and deliver energy efficiency.

Enbridge's Continuous Energy Improvement (CEI) initiative is focused on custom measures in the industrial segment. As part of ongoing modifications to this program, the industrial program will pursue greater targeting of small to mid-size operations and more flexibility in the incentives offered. As such, in 2012 Enbridge proposes to increase its custom incentive and expand its prescriptive offering to include more measures. Greater segment-focused marketing activities aimed at the mid-size facilities will augment the traditional marketing efforts for larger customers.

Figure 2 presents the commercial and industrial sector divisions of Enbridge custom projects in 2011. The number of projects in the commercial sector is more than six times the number of the projects in the industrial sector, but the average commercial sector project is only about one third the size of the average industrial sector project.

Figure 2. Enbridge 2011 Custom Projects Overview

Enbridge Custom Sector	# of Custom Projects	Gas Savings	% of Custom Portfolio
Commercial	780	37,470,116	68.2%
Industrial	127	17,482,847	31.8%
Total	**907**	**54,952,963**	**100%**

Source: Enbridge Gas Distribution Company

There are important differences in the Union and Enbridge custom programs. One difference is the average size of project. The average Enbridge commercial project is about 48K therms compared to about 99K therms for the Union C/I market projects. The average Enbridge industrial project is about 138K therms compared to the Union T1/R100 industrial projects, which average about 493K therms. In general terms, Enbridge's programs serve a market more dominated by commercial customers with smaller average project sizes, while Union's programs generally serve a market with more industrial customers, which results in larger projects in terms of savings. These factors need to be taken into account in an efficient sample design.

4. Analysis of Sampling Methodologies in Selected Jurisdictions

This section presents the findings from a review of sampling methodologies used in the evaluation of custom project programs in North America, including those described in annual evaluation reports of selected utilities as well as methodologies contained within evaluation protocols. The reviewed methodologies are all contained within publicly available documents. Because the reviewed documents contain varying degrees of detail and explanation, the Navigant Consulting, Inc. (Navigant) team applied its best interpretation of these documents to synthesize the available information in a consistent manner.

4.1 Summary of Jurisdictions Reviewed

The analysis of the reviewed methodologies accounts for factors such as fuel type, customer segment, and program design factors that might influence the design of samples for realization rate analyses.

Seventeen documents[8] were reviewed covering 12 unique jurisdictions in North America listed below:

- Illinois (Chicago) – Commonwealth Edison Company[9]
- Michigan (Detroit) – DTE Energy[10]
- Massachusetts – Massachusetts Energy Efficiency Advisory Council[11] covering NSTAR, National Grid, and Western Massachusetts Electric Company
- New Mexico – El Paso Electric Company,[12] New Mexico Gas Company,[13] and Public Service Company of New Mexico[14]
- Pennsylvania (Philadelphia) – PECO Energy Company[15,16]
- Ohio – AEP Ohio[17]

[8] Not counting the review of methodologies used by Union and Enbridge in prior evaluation cycles.

[9] "Evaluation Report: Smart Ideas for Your Business Custom Program." (Program Cycle 2010-2011.) Commonwealth Edison Company. Prepared by Navigant Consulting, Incorporated. May 16, 2012.

[10] "Reconciliation Report for DTE Energy's 2010 Energy Optimization Programs." DTE Energy Company. Prepared by Opinion Dynamics Corporation. April 15, 2011.

[11] "Impact Evaluation of 2008 and 2009 Custom CDA Installations." Massachusetts Energy Efficiency Advisory Council. Prepared by KEMA and SBW Consulting Incorporated. June 7, 2011.

[12] "Evaluation of 2011 DSM Portfolio." El Paso Electric Company. Prepared by ADM Associates Incorporated. May 2012.

[13] "Evaluation of 2011 DSM Portfolio." New Mexico Gas Company. Prepared by ADM Associates Incorporated. June 2012.

[14] "Evaluation of 2011 DSM & Demand Response Portfolio. "Public Service Company of New Mexico. Prepared by ADM Associates Incorporated. March 2012.

[15] "Annual Report to the Pennsylvania Public Utility Commission for the Period June 2010 through May 2011." PECO Energy Company. Prepared by Navigant Consulting. November 15, 2011.

[16] "Audit Plan and Evaluation Framework for Pennsylvania Act 129 Energy Efficiency and Conservation Programs. "Pennsylvania Public Utility Commission. Prepared by the PA Statewide Evaluation Team. November 4, 2011.

[17] "Program Year 2011 Evaluation Report: Business Custom Program." AEP Ohio. Prepared by Navigant Consulting, Incorporated. May 10, 2012.

- Maryland – EmPOWER Maryland[18] covering Baltimore Gas & Electric, Potomac Electric Power Company, Delmarva Power, Southern Maryland Electric Cooperative, and Potomac Edison
- California – California Public Utilities Commission,[19,20,21]covering Pacific Gas & Electric, Southern California Edison, Southern California Gas, and San Diego Gas & Electric
- Vermont – Vermont Department of Public Service[22] covering Efficiency Vermont and Burlington Electric Department
- PJM Interconnection – covering participating utilities in the Midwest and Eastern U.S.[23]
- U.S. Federally Owned Facilities – U.S. Department of Energy[24]
- International Performance Measurement and Verification Protocol (IPMVP) – Efficiency Evaluation Organization[25]

Figure 3 provides a high-level summary comparing the reviewed studies and Appendix C presents more detail on methods used in selected jurisdictions.

4.2 Key Findings – Review of Methods Used in Selected Jurisdictions

Commercial and industrial programs across North America range in type and size, and they frequently use inconsistent nomenclature. It is common to see custom C&I programs separated from prescriptive programs; however, some utilities do combine custom and prescriptive measures into a single program. Stratification approaches and confidence and precision targets are determined differently, depending on each utility's regulatory requirements and program organization.

Many publicly available evaluation reports tend not to describe sampling methodologies in much detail. These reports focus more on reporting evaluation results rather than describing methods used. Certain attributes of the sampling methodologies can be deduced from the reports, but explicit detail on the sampling approach ranges from little to none. The Navigant team applied its best interpretation in assessing utility evaluation reports.

---

[18]"EmPower Maryland 2011 Evaluation Report – Chapter 4: Commercial and Industrial Custom and Re-commissioning Programs." Baltimore Gas & Electric, Potomac Electric Power Company, Delmarva Power, Southern Maryland Electric Cooperative, and Potomac Edison. Prepared by Navigant Consulting, Incorporated.

[19]"Energy Efficiency Evaluation Report for the 2009 Bridge Funding Period." California Public Utilities Commission. January 2011.

[20]"The California Evaluation Framework." California Public Utilities Commission. Prepared by TecMarket Works. June 2004.

[21]"California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals." California Public Utilities Commission. Prepared by TecMarket Works. April 2006.

[22]"Verification of Efficiency Vermont's Energy Efficiency Portfolio for the ISO-NE Forward Capacity Market." Vermont Department of Public Service. Prepared by West Hill Energy and Computing Incorporated. July 29, 2010.

[23]"PJM Manual 18B: Energy Efficiency Measurement & Verification." PJM Forward Market Operations. March 1, 2010.

[24]"M&V Guidelines: Measurement and Verification for Federal Energy Projects Version 3." U.S. Department of Energy. Prepared by Nexant Incorporated. April 2008.

[25]"International Performance Measurement and Verification Protocol: Concepts for Determining Energy and Water Savings Volume 1." Efficiency Valuation Organization. January 2012.

NAVIGANT

Figure 3. Summary Comparison of Sample Methodologies in Selected Jurisdictions

N°	Service Territory or Jurisdiction	Organizations Reviewed	Year	Service Type	Timing	Precison Target	Stratify by Size	Stratify by Segment	Ratio Estimation
1	Illinois (Chicago)	Commonwealth Edison Company	2011	Electric	2-stage	90/08 (3yr utility program)	✓		✓
2	Michigan (Detroit)	DTE Energy	2010	Gas & Electric	1-stage	90/10 (utility program)		✓	✓
3	Massachusetts	Massachusetts Energy Efficiency Advisory Council (NSTAR, National Grid, Western Massachusetts Electric Company)	2009	Gas & Electric	1-stage	90/10 (statewide custom C&I)			✓
4	New Mexico	El Paso Electric Company, New Mexico Gas Company, Public Service Company of New Mexico	2011	Gas & Electric	1-stage	90/10 (utility total portfolio)	✓		✓
5	Pennsylvania (Philadelphia)	PECO Energy Company	2011	Gas & Electric	3-stage	85/15 (utility C&I total)	✓	✓	✓
6	Ohio	AEP Ohio	2011	Electric	2-stage	90/10 (utility program, RTO zone)	✓	✓	✓
7	Maryland	EmPower Maryland (Baltimore Gas & Electric, Potomac Electric Power Company, Delmarva Power, Southern Maryland Electric Cooperative, and Potomac Edison)	2011	Gas & Electric	1-stage	80/20 one-sided (utility program)	✓		✓
8	California	California Public Utilities Commission (Pacific Gas & Electric Company, San Diego Gas & Electric, Southern California Edison, Southern California Gas Company)	2009	Gas & Electric	flexible	90/10 (utility program)	✓	✓	✓
9	Vermont	Vermont Department of Public Service (Efficiency Vermont and Burlington Electric Department)	2010	Electric	2-stage	80/10 (utility portfolio)	✓	✓	✓
10	PJM Interconnection (Midwest & Eastern US)	PJM Interconnection	2010	Electric	flexible	90/10 one-sided (utility program, RTO zone)	✓	✓	✓
11	US Federal Facilities	US Department of Energy	2008	not applicable	flexible	not applicable		✓	
12	General International	Efficiency Valuation Organization (IPMVP)	2012	not applicable	flexible	not applicable		✓	

Source: Navigant review of previously cited documents in selected jurisdictions

Protocols for evaluating DSM projects in specific jurisdictions tend to provide a more detailed description of sampling methodologies used than the program evaluation reports. Protocols generally allow specific sampling options such as selecting between census, simple random sampling, and stratified sampling, as well as options for determining the appropriate basis for stratification. The reviewed protocols usually offer step-by-step processes for designing samples.

Meeting Precision Targets

Confidence and precision requirements vary widely across the reviewed methodologies. Both one-sided and two-sided confidence intervals are common. Confidence requirements range from 80% to 90%, and precision requirements ranged from 8% to 20%. These confidence and precision requirements frequently differ in the level at which they are applied, which could be for the program, the customer segment, the portfolio, or the transmission zone. One methodology[26] adheres to a relatively rigorous precision target of 90/08, but the target only applies to a 3-year term rather than annually.

On-site verification and evaluation is common industry practice for evaluating larger custom program impacts. There are cases where phone and engineering algorithm verifications have been used for custom programs in some years with more in-depth evaluation work performed in other years. Phone surveys are generally reserved for process evaluation and establishing free-ridership estimates. Phone surveys are less commonly used to estimate gross program impacts. The reviewed methodologies tend to contain a rather substantial description of the evaluation techniques used to estimate project savings, often describing in detail the engineering models applied and how parameters were measured and used. Several evaluation sample design methodologies apply more rigorous techniques or aim to achieve a census for large projects that represent a high concentration of savings in order to cost-effectively increase validity and accuracy of evaluation estimates at the project and program levels.[27,28]

Ratio estimation is used in nearly all of the reviewed methodologies and has now become a standard practice in the industry. Ratio estimation is a statistical technique whereby prior information from a tracking database—"tracked savings"—is employed to reduce the overall sample requirements. If stratification is used, the resulting precision is applied to the total based on applying the realization rate measured for each stratum.

An expected variance must be assumed to create an initial sample design. This assumption is made via an error ratio or coefficient of variation (CV). The CV is defined as the standard

[26] "Evaluation Report: Smart Ideas for Your Business Custom Program." (Program Cycle 2010-2011.) Commonwealth Edison Company. Prepared by Navigant Consulting, Incorporated. May 16, 2012.

[27] As a point of interest, the more rigorous evaluation approaches for selected large projects can, on occasion, produce a higher variance across the sample. This can produce the appearance of worsening sampling precision, but it is generally viewed as producing more appropriate levels of confidence and precision for the program.

[28] "EmPower Maryland 2011 Evaluation Report – Chapter 4: Commercial and Industrial Custom and Re-commissioning Programs." Prepared by Navigant Consulting, Inc.

deviation of the sample divided by the mean. In the case of ratio estimation, the CV should be based on the variance of project-specific realization rates rather than the variance of savings. Industry practice is to conservatively rely on historic evaluation results in selecting a CV for sample design. When historic data are not available, conservative assumptions are made, typically ranging from 0.5 to 1.0 depending on the expected homogeneity of the population.[29] Ratio estimation can sometimes reduce the CV to levels around 0.3; however, these levels represent "best outcomes" and should not be viewed as conservative when designing a sampling framework.

The reviewed methodologies more commonly apply Z-values[30,31] than T-values in determining sample precision. At larger sample sizes (i.e., greater than 30) the differences are insignificant. But for smaller samples, application of the Z-value fails to account for the limited degrees of freedom in the sample and can lead to overstating the confidence and precision achieved by the sample.

Use of the finite population correction (FPC) factor is not frequently discussed. However, the FPC has a valid statistical basis and should be used when evaluating smaller populations. Two of the reviewed methodologies[32,33] do not appear to use the FPC, and instead recommend a census if the calculated sample size approached or exceeded the population size. Any sample size calculation that exceeds the population is not taking into account the basic principles of sample design. This approach is not statistically valid and can lead to excessive evaluation costs. Although this topic is not frequently discussed, it is reasonable to assume that the FPC is applied whenever size-based sampling was used since application of the FPC is necessary to take advantage of the concentrations of savings in large projects.

Use of Stratification

The reviewed methodologies applied stratification in the sample design when population sizes were not sufficiently small to achieve a census. Stratification approaches vary across the reviewed methodologies and appear to be customized to fit each utility's program structure, number of projects, sizes of projects, regulatory requirements, and stakeholder concerns.

The review yielded two common approaches for stratifying based on size. The first approach defines the large stratum based on very large projects in the population. Sometimes a census is

[29] "PJM Manual 18B: Energy Efficiency Measurement & Verification." PJM Forward Market Operations. March 1, 2010. (See page 30)

[30] "Audit Plan and Evaluation Framework for Pennsylvania Act 129 Energy Efficiency and Conservation Programs." Pennsylvania Public Utility Commission. Prepared by the PA Statewide Evaluation Team. November 4, 2011.

[31] "The California Evaluation Framework." California Public Utilities Commission. Prepared by TecMarket Works. June 2004.

[32] "The California Evaluation Framework." California Public Utilities Commission. Prepared by TecMarket Works. June 2004. (See page 337)

[33] "Audit Plan and Evaluation Framework for Pennsylvania Act 129 Energy Efficiency and Conservation Programs." Pennsylvania Public Utility Commission .Prepared by the PA Statewide Evaluation Team. November 4, 2011. (see page 75)

sought when the very large stratum contains only a few projects. The second approach divides the population into strata of roughly equal contribution to total savings.[34] In some cases, this approach seemed to follow textbook examples rather than examining the program projects to see if alternate approaches to stratification could be designed to increase precision. Simply dividing the population into three roughly equal strata may overlook more appropriate stratification designs that could yield higher precision and confidence. This approach is more applicable when project size declines smoothly from large to small projects. Some of the reviewed methodologies apply more rigorous evaluation and measurement approaches to projects in the large stratum or for strata with highly heterogeneous populations in a cost-efficient effort to improve accuracy.

Many of the reviewed methodologies stratify by segment instead of or in addition to stratifying by size. Segments used for stratification included market sector (e.g., education, multi-family, manufacturing, and other customer-type segments), geography, and project types (space heating, water heating, or industrial process). Stratification by segment can be used to increase precision for a given sample size as well as make the sample more representative of the population.

Sample Staging

Schedule requirements for reporting often necessitate a rolling sample or staged approach to sampling in order to begin evaluation efforts early enough to complete the evaluation tasks in time to report results on schedule. About half of the reviewed methodologies implement staged sampling. Most of the methodologies do not require reporting intermediate results, but rather focus only on the final population results.[35]

A two-stage approach is most common[36,37,38] where a stage one sample is drawn based on either the first two or first three quarters of the year. Single-stage sampling and three-stage sampling also occur in the reviewed methodologies. Details on the rationale underlying the calendar periods for the different stages, and the allocation of sample to the different stages, were generally not explicitly stated. In general, approaches were based on "reasonable judgment" by the evaluators.

---

[34] "Program Year 2011 Evaluation Report: Business Custom Program." AEP Ohio. Prepared by Navigant Consulting, Incorporated. May 10, 2012. (See appendix J, page 33)

[35] Pennsylvania has a slight exception. Reporting quarterly results is required by Act 129. Although quarterly reporting has been interpreted as applying to unverified results, verified results are reported for the full year.

[36] "Evaluation Report: Smart Ideas for Your Business Custom Program." (Program Cycle 2010-2011.) Commonwealth Edison Company. Prepared by Navigant Consulting, Incorporated. May 16, 2012.

[37] "Program Year 2011 Evaluation Report: Business Custom Program." AEP Ohio. Prepared by Navigant Consulting, Incorporated. May 10, 2012. (See appendix J, page 33)

[38] "Verification of Efficiency Vermont's Energy Efficiency Portfolio for the ISO-NE Forward Capacity Market." Vermont Department of Public Service. Prepared by West Hill Energy and Computing Incorporated. July 29, 2010.

Gas & Electric Service

Major differences in evaluating savings between electric and gas utilities were not found. Differences in evaluation methods are more likely based on program size and number of years evaluating and reporting program savings. Most jurisdictions count both electric and gas savings for custom C&I measures regardless of whether the administrating utility supplies both fuel types.

Bias in Results

Industry best practices prescribe a demonstration of effort to control for common sources of bias. Once a population of projects exists, the goal of the sample design is to estimate the gross savings resulting from that population.[39] The principal concern about bias is that certain elements of the population may be over- or underrepresented in the sample. Stratification is a good approach for reducing this potential bias. Bias can also result from non-random sample selection. Finally, bias can be introduced into the analysis by anomalous observations in the sample that for some reason are unique and not representative of other members of the population. If anomalous observations are also "influential" observations, then corrective action may be necessary to provide accurate information from the realization rate calculation, and the accompanying calculations of precision and confidence. The California Evaluation Framework notes:[40,41]

> *[If] there is substantial bias, perhaps due to self-selection, non-response, deliberate substitution of sample projects, or measurement bias, then the methods presented here can be seriously misleading. For example it is misleading and counterproductive to report that the average savings has been estimated with a relative precision of 10% at the 90% level of confidence if there is a serious risk that the results might be in error by 25% due to bias.* (p. 327)

The reviewed methodologies contain little description of efforts made to minimize bias. Additionally, there is little discussion on the composition of the sample, treatment of outliers, sample replacements, missing data points, or other sample adjustments. These discussions could be addressed in project memos rather than expanding what is often a lengthy final evaluation report. However, this is an area where standard industry practice may not be on par with evaluation practices in other fields. It is not clear whether this deficiency is related only to reporting or if it reflects limitations on current evaluation practice.

---

[39] Issues such as self-selection bias in recruiting program participation are not an issue for sample designs whose purpose is to estimate the gross savings from those that did participate in the program. Once the frame of participant projects is determined, the biases of concern are typically based on ensuring random samples, ensuring representativeness, addressing extreme values, and using appropriate calculations consistent with the sample cases to produced unbiased estimates of the population parameters.

[40] "The California Evaluation Framework." California Public Utilities Commission. Prepared by TecMarket Works. June 2004.

[41] The California Evaluation Framework contains a substantive discussion on accuracy and bias in chapter 12.

5. Recommended Sample Design Methodology

This section describes the recommended sample design methodology for DSM programs for Union and Enbridge. Sections 5.1–5.3 describe the key attributes of the recommended methodology and offer support for their use in evaluating Union and Enbridge custom programs. Section 5.4 presents steps for appropriate sample designs and sample selection. Sections 5.5–5.6 present examples for Union and Enbridge illustrating how the sample methodology might be implemented using representative tracking data.

Ratio estimation has become standard practice for the evaluation of large C&I programs, as it leverages information available on the population of projects with the sample. The sample design approaches discussed in this section are constructed to make full use of the ability to leverage sample data in combination with information on the population from the project tracking database. This is important given the relatively high cost of rigorously evaluating custom C&I projects. Ratio estimation has become a common industry practice in evaluation since it leverages information on the population to better interpret information from the sample. Stratification has also become a common industry practice, although its application varies, and its application may not result in strata that enhance the efficiency of the sample design. The methods presented in this section are aligned with these basic concepts of leveraging information to get the most out of the analysis.

The level of specification for sampling protocols observed in jurisdictions across North America ranges widely. An overly specified methodology may lead to incompatibilities in future evaluation efforts as the composition, participation, and distribution of impacts evolve. However, an overly general methodology may lead to sample designs that do not meet Union and Enbridge's confidence and precision requirements with cost-efficient methods. The recommended sample design methodology is intended to strike a balance between flexibility and specification to allow Union and Enbridge to best meet their evaluation needs now and in future program years.

5.1 Stratification

Stratification is recommended in designing samples for evaluating custom C&I programs. Stratification is the practice of disaggregating the population into sub-groups based on some criteria. Strata should be defined such that the strata sample frames are mutually exclusive (i.e., no overlap) and exhaustive (i.e., strata sample frames combine to represent the appropriate population sample frame). There are three generally accepted reasons to use stratification:

1. Sample Efficiency: To reduce the required sample size needed to achieve confidence and precision targets on an estimate. There are two common stratification practices that can increase sample efficiency:

- Stratifying by project size may reduce the overall number of required samples by taking advantage of the concentrations of savings when relatively few projects contribute to a large fraction of total impacts. This is most commonly seen in C&I evaluations, and the majority of reviewed methodologies apply this approach.
- Stratifying based on qualitative segments (e.g., project type or customer segment) can reduce the effective variance compared to combining the segments in a single stratum when segments of a population produce different results. For example, if the project-level realization rate (RR) is expected to average 0.9 for lighting projects and 0.8 for heating, ventilating, and air conditioning (HVAC) projects, then the variance of these segments combined will usually be greater than their individual variances. Separating lighting from HVAC would then allow smaller sample sizes to meet the required precision criteria for total combined savings.

  Stratification design must reduce the effective sample variance in order to produce gains in precision. The simple rule is that projects within a sample should have a smaller variance within the strata than across strata. Lohr notes:[42,]

  > *Observations within many strata tend to be more homogeneous than observations in the population as a whole, and the reduction in variance in the individual strata often leads to a reduced variance for the population estimate.* (p. 77)

- Stratification cannot make the problem worse (i.e., decrease precision). As a result, it is strongly recommended.

2. Segment Results Required: To ensure sufficient sample sizes that can answer questions pertaining to certain segments of the total population**.** For example, if stakeholders or interveners require results specifically for HVAC-related projects in order to improve program implementation in subsequent years, then creating strata for HVAC projects and establishing a minimum precision requirement for those strata would help ensure that sufficient data are collected to understand HVAC projects.
3. Reduced Potential for Bias by Improving the Representativeness of the Sample: For many evaluators, this is the most important reason for stratification as part of sample design. Stratification helps ensure that the sample appropriately represents the population. Since simple random sampling allows for the possibility of under-sampling certain segments, stratification can help ensure that the sample drawn provides the appropriate sample size for each segment. For example, stratifying by project type can ensure that each major project category is appropriately represented in the sample by explicitly drawing samples for each project type. Other frequently used dimensions for stratification include customer segments and site geographies. Representativeness quotas are sometimes used instead of strata to ensure representativeness.

[42] Lohr, S. L., "Sampling: Design and Analysis," Second Edition, 2010.

The specific stratification approach will depend on evaluation of the population data. If the distribution of project savings for a program is relatively tight[43] and there is not an easily delineated group of large projects, then stratification by project size alone may not produce sampling efficiencies. However, if the distribution of project savings is wide or there is clear group of large projects, then stratifying by project size will likely produce sampling efficiencies.

It is important to note that when sample observations are collected based on a stratified sample design, the strata weights must be applied in the estimation of the population realization rate.

The general rule for stratification is to attempt to select strata that have smaller variance within the strata than between strata. Stratifying by segment may also be appropriate when realization rates are expected to vary by segment. Judgment should be applied to segment the population on the basis of mechanisms that lead to different realization rates, rather than simply using common predefined segments used in program administration. For example, if steam projects are expected to have a different realization rate than other project types—or even more widely varied realization rates across steam projects—then a potentially useful segmentation may be by steam projects vs. other non-steam projects. It is not necessary to segment by every major project category to achieve the desired sampling efficiency, only those where this effect is believed to be sizeable and where stratification may also help increase the representativeness of the final sample across important technology categories.

5.2 Ratio Estimation

The application of a ratio estimation approach is recommended. Ratio estimation is the statistical technique whereby the *accuracy* of "prior" tracked estimates is applied from the sample rather than directly applying the *absolute* estimates of the sample. For DSM evaluation efforts, the sample estimator is the realization rate for each stratum rather than the sampled savings for each stratum. Ratio estimation is often used to increase the precision of estimated means and totals. It is motivated by the desire to use information about a known auxiliary quantity (i.e., tracked savings) to obtain a more accurate estimator of the population total or mean (i.e., verified savings). When applying ratio estimation within a stratified population, the separate ratio estimator approach should be used where strata are defined and analyzed before combining strata.[44]

Ratio estimation would not be possible without initial savings estimates for the population. This technique relies on establishing the variance based on the errors between the savings predicted by the stratum average realization rates for each project and the actual savings measured for each project. Ratio estimation effectively develops verified savings estimates based on measuring the accuracy of the tracked savings. Therefore, it is necessary to ensure that the tracked savings in the tracking database represent the best possible estimate based on the available information.

[43] A "tight" project savings distribution is generally considered to be within a single order of magnitude. Size-based stratification should be considered when the distribution of savings spans multiple orders of magnitude.

[44] Lohr, S. L., "Sampling: Design and Analysis," Second Edition, 2010. (Section 4.5)

5.3 Sample Staging

A rolling sampling approach comprised of two sample draws (a two-stage sample approach) is recommended to ensure that spring reporting requirements can be met. Reporting schedules often do not provide sufficient time to design and evaluate a sample following the completion of the project year. This type of schedule constraint frequently occurred in the jurisdiction reviewed in Section 4. Sample staging can allow evaluation efforts to begin earlier on a preliminary sub-sample of projects completed early in the program year. Thus, staging can reduce the evaluation workload required between the end of the program year and the reporting deadline.

A two-stage sample is recommended, where the first stage takes a sample draw from projects completed in the first three quarters of the program year, and the second sample draw adds in projects completed in the fourth quarter.

The sample design for the first stage should estimate or extrapolate the numbers of projects in each stratum to the values expected at the end of the year.[45,46] Sample sizes should be determined for this preliminary sample frame as an indication of the final population. While judgment is needed to determine how much of the expected overall sample is drawn in the first stage, it is unlikely that the first stage sample would fully require three-quarters of the calculated sample sizes.[47] In general, practical considerations would support a lower split of the planned sample between the first and second stages. This would allow for a sample that adequately represents the year-end projects.

Union's and Enbridge's projects tend to come online more heavily in the fourth quarter, with roughly half to three-quarters (depending on which program) of projects completing in the last quarter. This would imply that a 50-50 split between sample stages would be reasonable, given constraints related to the calendar time needed to set up and conduct the verification studies. However, if the timing allows, Union and Enbridge might consider placing more of the sample into the fourth quarter when savings from projects completed in the fourth quarter are expected to contribute more than half of program savings. This recommendation is a compromise between the time and resources needed to perform the number of site verifications, and the need to meet program reporting deadlines. It simply is not possible for the utilities to wait until information on that year's full population of projects becomes available and then draw the sample and complete the site verifications while still meeting the program reporting deadlines.

---

[45] This step is important because it will reduce the effect of finite population correction that could otherwise lead to underestimating the required sample sizes.

[46] If the final quarter of the program year is known to have very large projects in disproportion to the first three quarters, the strata weighting may be adjusted to account for this information.

[47] The sample sizes may be further reduced slightly to allow for the possibility that the assumed CV is overly conservative. If upon evaluation of the first stage, the assumed CV was not overly conservative, then additional samples may be added in the second stage.

This rolling sample or two-stage approach is often used in program evaluation (see Section 4 above) to meet timely reporting deadlines.

The sample design for the second stage should consider the population of the program year in its entirety. Sample sizes should be determined for the entire population. The first stage sample is intended to fulfill about half of the overall sample. The second stage is intended to fulfill the remainder of the sample and should be selected from projects completed in the fourth quarter.[48] If analysis of the first stage sample observations indicates insufficient sample sizes, then the first stage may be reinforced in the second stage with additional projects selected at random from the full program year population. An analysis of sample data should investigate whether differences between sample stages are significant and adjustments are needed. Again, the goal is to produce good information for making decisions regarding the custom programs for both the utilities and stakeholders. Some judgment is needed in implementing this rolling two-stage sample selection approach.

5.4 Recommended Sample Design Process—Seven Steps

The sample study should be designed to estimate the impacts of the population of projects in each program year. At the time of this report, gross *cumulative* (i.e. lifetime) gas savings measured in cubic meters (m^3) is the primary impact to be studied and should serve as the basis of the sample design.[49] The sampling and the application of population-wide realization rates should all be performed using gross cumulative savings.[50] The recommended sample design methodology contains the following steps:

Step 1: Review project tracking database for accuracy and quality.
Prior to any stratification or sampling, large gains can be made in the resulting analysis and precision by reviewing the estimates in the tracking database and making sure that the best possible initial project-based engineering estimates are contained in the tracking database. It is also important to make sure that appropriate contact information is contained in the files to avoid having to replace drawn sample projects with supplemental projects held in reserve. One of the most cost-effective ways to enhance the precision and confidence in the evaluation results is to make the appropriate investment in the tracking database. A tracking database that is accurate will typically reduce the costs of the evaluation, yield project realization rates that are closer to one, and have a smaller variance across the project realization rates. Many utilities do a

[48] Although this approach is intended to achieve roughly equal proportions of projects for each quarter, disproportions by quarter should not be viewed as causing notable bias. Accordingly, if the first stage produces a small number of projects in excess of what is required in the second stage, these extra projects may be counted toward meeting the fourth quarter sample size requirements.

[49] This is a new basis for custom C&I evaluation studies beginning in program year 2012. The Technical Evaluation Committee may decide to change this basis in future years.

[50] Ultimately, adjusted gross savings can be converted to adjusted net savings (i.e. by applying a program net-to-gross ratio to the adjusted program gross savings). However, that would occur outside of (i.e. after) the application of the sampling work discussed in this report.

second check of the tracking database prior to the sample design and sample selection.

Identifying unique projects in the tracking database can help avoid outlier problems later in the analysis. Examples of unique projects may be those with the only instance of a certain efficient technology installed or even those with technologies whose impacts are difficult to predict. These unique projects may be treated separately from the primary population to produce more efficient samples for the vast majority of the population. Identification of unique projects can also help ensure the representativeness of the selected sample and help eliminate problems in the interpretation of the analysis such as bias in the realization rate.

Step 2: Evaluate the population and define strata.
Examine the population for ways to leverage the sample design to improve efficiencies in meeting target confidence and precision levels. This includes three activities:

- *Exclusion of extremely small projects* – Ratio estimation weights project realization rates according to project savings. Very small projects typically exert only negligible influence on estimates of the total realization rate, the total savings, and the total achieved precision. For many very small projects, a 100% difference in realized savings would produce a negligible impact on the total estimates. The cost of evaluating the impacts of these small projects exceeds the value of the information obtained from them. Additionally, including projects that contribute only small fractions of a percent to program savings in the sample frame might result in the random selection of projects that includes a disproportionate number of these very small projects, which could reduce the accuracy with which the overall realization rate is estimated for a given sample size and reduce the overall representativeness of the sample. It is therefore considered reasonable to exclude the very small projects (i.e., representing up to 5% of the total program savings as appropriate) from the sample frame. The savings of the population of very small projects may be adjusted by an appropriate realization rate[51] and added to the program savings total.
- *Identification of project size strata bounds* – Efficiencies can be gained by stratifying by project size when the distribution of project savings is wide or there is a clear group of large projects. Sorting the projects by savings size can allow easy identification of discontinuities in the project size distribution. If it is unclear whether natural project size groupings exist; visualization of the project savings in a histogram should provide a clearer indication. Typically, strata are set such that program savings within a stratum fall within an order of magnitude.[52] Set strata bounds first based on natural breaks in the distribution that result in easily delineated groupings. If natural groupings do not exist,

[51] If the remaining population is stratified by size, then the average small stratum realization rate should be applied. Otherwise the population total realization rate should be applied. However, the savings accounted for by these projects is so small that alternative assumptions should not affect the overall program savings estimates. Some applications simply use a realization rate of 1.0 for these very small projects.

[52] One rule of thumb is to keep the expected coefficient of variation of project savings to less than 1.0 within a stratum.

other approaches may be used such as stratifying into strata of roughly equal total savings. The number of size-based strata typically ranges from two to four, with three most commonly applied for C&I program evaluations.

- *Identification of categorical characteristic strata bounds* – Efficiencies can be gained by defining strata along categorical qualities such that the coefficient of variation of project realization rates for each stratum is lower than the resulting CV of the aggregated group without the categorical strata. This basis for stratifying may be applicable when a certain segment of the project population is expected to have different or more variable realization rates than the rest of the population. Units that are generally more alike should be grouped together in a stratum. For commercial projects, strata could be defined by building type (e.g., schools, office building, and multi-family). Similar buildings could be expected to have a lower variance in the estimated realization rate across sites (i.e., within the stratum) than when combined with other building types. Although categorical strata bounds are frequently applied in many DSM studies, they are not mandatory and should be prudently applied.

The sample designer may be required to make trade-offs between stratification approaches. Defining the appropriate strata is often the most important part of sample design; however, it requires data analysis skills, subject matter expertise on the project types, and knowledge of program administration and participation issues.

Step 3: Estimate an appropriate variance for each stratum.
In ratio estimation, the variance considered is that of the residuals on the stratum average realization rate rather than the variance of the verified savings. Accordingly, a CV or error ratio should be based on the assumed distribution of individual realization rates for the population of projects in each stratum.

The CVs should be based on the un-weighted[53] realization rates historic sample data, when such data are available. Any changes in program composition, administration, or participation from the previous year will decrease the validity of applying prior year CVs, and the assumed CVs should be adjusted upward by 0.1-0.2 to prevent under-sampling. It is not recommended to apply a coefficient of variation less than 0.30, in order to ensure sample sizes sufficient for robust results and to allow for increasing variances that may result from evolving measurement approaches and program participation.

A two-staged sample provides an opportunity to adjust the assumed CVs in the second stage to incorporate the sample data already observed in the first stage. The observed CVs in the first stage should still be slightly adjusted upward to account for variance and size unknowns in the second stage sample.

[53] The realization rates are un-weighted rather than weighted because it is assumed that any correlation between the size of a project in a stratum and its realization rate is coincidental (especially in small sample sizes). So, applying the historic correlation could result in under-sampling or over-sampling in subsequent program evaluation efforts.

A CV of 0.5 may be assumed when historic data are not available. This is a standard industry assumption and is generally conservative in ratio estimation if the population tracked savings in the tracking database are reasonably accurate. However, custom projects with poor tracking database estimates may produces CVs as large as 1.0. It is not uncommon to observe program CV's lowering over time as programs mature and tracking estimates improve. CVs can also increase if more rigorous and precise methods are used to evaluate project savings; however, this should not be viewed as a negative since rigorous methods create a more accurate understanding of project and program results.

Step 4: Allocate observations to each stratum.
The overall sample should be designed to achieve 10% precision at a 90% one-sided confidence level (i.e., 90/10 one-sided).[54, 55] This confidence and precision target is meant to be used for each custom program in each year. If changes are made to this target, these changes can be addressed in the sample size calculations and do not necessarily warrant changes in the recommended methodology. Appendix A and Figure 19 provide additional explanation and illustration for the 90/10 one-sided confidence interval and the other reporting confidence intervals.

Allocating the sample across strata to achieve target confidence and precision is not a simple exercise and can often require an iterative approach. Proportional sampling is one technique that is often applied, where the total sample size is calculated for the population and subsequently allocated to strata in proportion to some characteristic such as savings. Proportional sampling, however, fails to realize the efficiencies gained from stratifying and very frequently results in over-sampling. Lohr notes:[56]

> *If the variances are more or less equal across all the strata, proportional allocation is probably the best allocation for increasing precision. In cases where the variances vary greatly [across strata], optimal allocation can result in lower costs. In practice, when we are sampling units of different sizes, the larger units are likely to be pre variable than the smaller units [in absolute terms] and we would like to sample them with a higher fraction.*[57]

The California Evaluation Framework notes the skills required:

[54] Based on October 25, 2012 Technical Evaluation Committee decision, the sample design should be based on a 90/10 one-sided confidence interval. Reporting of achieved confidence and precision should present the precision achieved for three confidence intervals: 90% one-sided on the lower bound, 90% one-sided on the upper bound, and 90% two-sided intervals. Appendix A provides additional explanation and illustrative examples for these reporting confidence intervals.

[55] This target may be inferentially interpreted as the intent to ensure that there is a 90% likelihood that the actual savings of the program population exceeds 90% of the sample estimate of program population savings.

[56] Lohr, S. L., "Sampling: Design and Analysis," Second Edition.2010. (Section 3.4.2 discusses optimal allocation)

[57] Lohr, S. L., "Sampling: Design and Analysis," Second Edition.2010. (Section 3.4.2 discusses optimal allocation in more detail – p. 87.)

> *Stratified ratio estimation is somewhat more complex [than simple random sampling]…it probably still requires someone to have basic training and/or experience in statistics to ensure that it is understood and applied correctly.*[58]

Given the judgment needed to develop a sample design, it is important to test the robustness of the design by simulating different scenarios. Assessing several alternative allocations of the sample across strata can usually improve sample efficiency.

Step 5: Determine criteria for assessing sample representativeness. (optional)
There are often categorical characteristics of the population that are not used in defining strata but are still desired to ensure a reasonably representative sample.[59] For example, market segment may not have been used in defining strata; however, a random sample that fails to include certain major market segments would not be viewed as a representative sample. You could establish new strata for these factors; however, it is expected that a random draw will be representative across these factors and there is a benefit for a simple stratification design.

To address this, some criteria can be defined prior to randomly selecting a sample, which can be used to assess the representativeness of the sample. Criteria should be established only for the most important characteristics, and they should only be set for high-level characteristics that, if not met, would represent an extreme sample that would not representative of the population. Failure to meet the criteria will result in discarding the full original sample and selecting an alternate full sample. Criteria can be established only for the total population or specific strata as appropriate (See example in Section 5.5). Selection of a sample that does not meet representativeness criteria should be a rare occurrence. This approach is only meant to mitigate the possibility that a randomly selected sample might result in highly inaccurate statements about the entire population. The necessity to discard the original sample should not occur in most program years.

Step 6: Select a random sample.
The sample for each stratum should be selected at random from a uniform distribution. This provides an equal opportunity for each project within a stratum to be selected.[60] This can be accomplished in Microsoft Excel using the RAND() function[61] to assign a random number between 0 and 1 to each project in a stratum. The projects should be sorted within each stratum

[58] "The California Evaluation Framework." California Public Utilities Commission. Prepared by TecMarket Works. June 2004, p. 316.

[59] These criteria are not intended to be overly restrictive in selecting a sample. Rather, they are intended to prevent the unlikely but possible case where extreme over-representation or under-representation of certain project characteristics occurs in the sample.

[60] Sampling from a savings-weighted distribution can also be valid, but it is not recommended here since size-based strata are already employed.

[61] Note that the RAND() function will continue to generate a new set of random numbers each time a cell is updated. To prevent this, the values of the RAND() function can be copied and pasted (i.e., "paste values") into a separate column.

based on the random number assigned to it, and the projects with the highest random number should be selected for the sample until the target stratum sample size is reached.

The selected sample should be analyzed and documented. If criteria are set to assess the representativeness, the selected sample should be analyzed against these criteria at this point. If the sample does not meet the criteria for representativeness, then the full population sample should be discarded and a new sample should be selected.

Recruiting the full selected sample is often not achievable since some program participants may not respond or refuse to participate in the sample. Even when agreement to participate in evaluation activities is required to participate in the program, full recruitment of the selected sample can often not be achieved. Therefore, a set of potential replacement projects may be provided to recruiters to fill in for non-recruited participants.

Potential replacements should be selected from the same random number list of the population from which the original sample was selected. Replacements should be selected in priority of assigned random number until full recruitment is achieved. The full population of a stratum should not be provided to recruiters, whose incentives are not usually aligned to follow the random prioritization of the sample, unless the full sample size is not expected to be achieved.

Step 7: Recruit the sample.
Recruitment of each stratum sample can begin once the sample has been selected and assessed. Recruitment typically occurs over the phone, and may or may not involve scheduling of the on-site evaluation visit. Ensuring the accuracy and completeness of contact information in the tracking database can streamline the recruitment task.

The list of potential replacements may be initially withheld from recruiters to ensure that the originally selected sample projects are pursued fully before being replaced by alternate projects. This can help reduce the possibility for non-response bias in the sample. The California Evaluation Framework notes:[62]

> *It is very important to use the backup sample correctly. The most efficient way to recruit a sample of the desired size may appear to be to contact both the primary and backup sample at once and to schedule those sites that are first to respond and agree. But this is generally not sound practice since this approach ensures that the response will be no better than 50%, assuming that the backup sample size is equal to the primary sample size. Instead, the initial recruiting effort should be limited to the primary sample. A backup should be used only if a primary sample site is impossible to contact or refuses to participate.* (p. 350)

[62] "The California Evaluation Framework." California Public Utilities Commission. Prepared by TecMarket Works. June 2004.

A full effort should be made to recruit the original sample before resorting to replacements, and the same effort should be made to recruit each replacement before moving on to the next.

5.5 *Example Implementation of Sample Design Methodology (Union)*

This section demonstrates how the sample design methodology might be implemented for an example set of Union program data. The data used for this example has been randomized and does not indicate historic program achievements that have undergone regulatory review in prior years. The data for this example is intended to be representative of a typical program year and are used in this example for illustrative purposes only. This example is for reference and does not preclude the judgment needed to understand and address the idiosyncrasies of actual program data.

This example applies the seven steps of the sample design process presented in Section 5.4 above.

Step 1 reviews the project tracking database for accuracy and quality. Of particular emphasis is a check on the processes used to produce the initial estimates for savings contained in the database and the contact information. This step is usually undertaken by the utility and is done to provide the third-party evaluator with the best information possible. As mentioned above, a more accurate tracking database will make it more likely that confidence and precision targets will be met. This example assumes that the tracking database has been reviewed.

Step 2 evaluates the population and defines strata. Gross *cumulative* gas savings measured in cubic meters (m^3) is the primary impact to be studied and should serve as the basis of the sample design. Figure 4 and Figure 5 show representative project distributions of savings[63] for Union's T1/R100 and C/I programs, respectively. Analyzing the distribution of project sizes indicates that size-based stratification should produce sampling efficiencies. Other categorical bases for stratification are not chosen for this example, although Union may consider isolating new technologies into a unique stratum for future evaluation efforts.

[63] The initial manual produced in November, 2012 used net gas savings in the examples. In this revised report, the example analyses are performed on cumulative gross savings values to correctly illustrate how the sampling and the application of population-wide realization rates for the utilities should be performed in current sampling efforts.

Figure 4. Illustrative Distribution of Savings for Union'sT1/R100 Projects

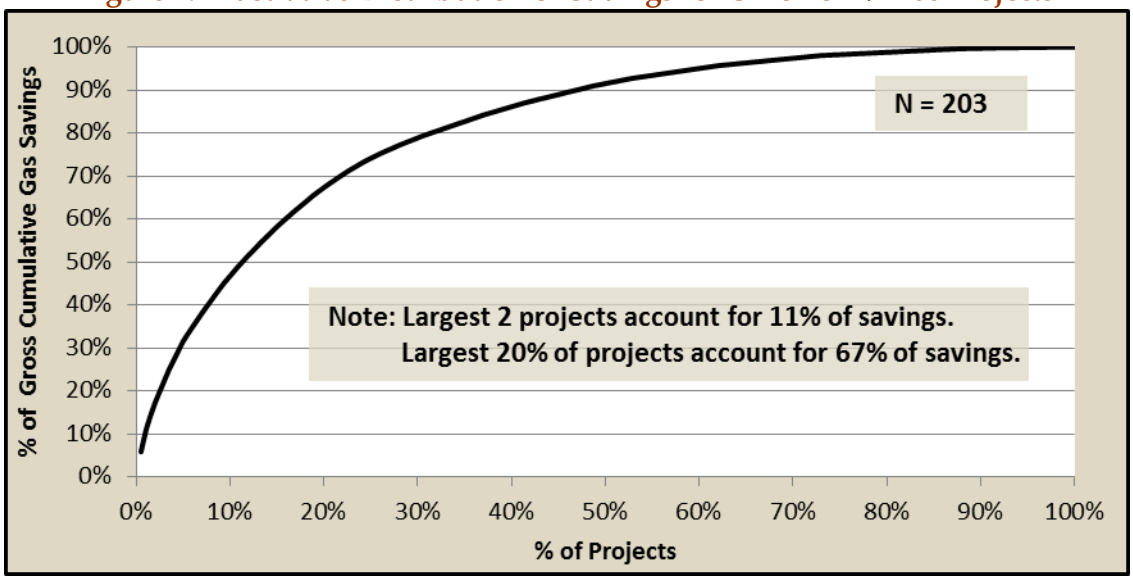
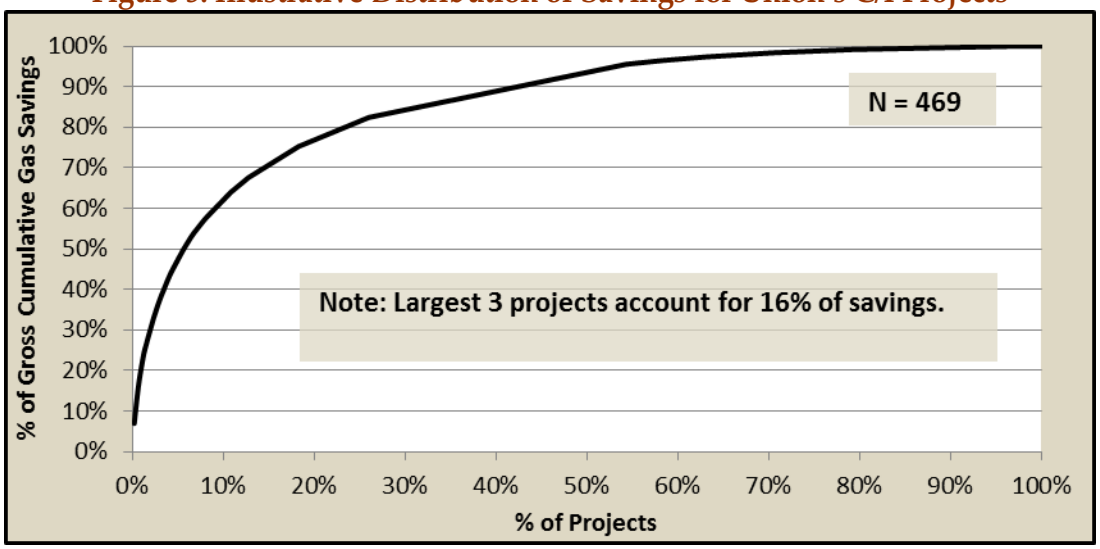


Figure 5. Illustrative Distribution of Savings for Union's C/I Projects



The sensitivity to sample sizes is investigated to determine appropriate savings thresholds for strata bounds. Figure 6 and Figure 7 show illustrative strata boundaries for Union's T1/R100 and C/I programs, respectively.

Figure 6. Illustrative Strata Boundaries for Union's T1/R100 Projects

Stratum Size	Lower Threshold of Cumulative Gross Gas Savings (m^3)	Projects	Savings Represented (%)
Large	50,000,000	10	31.4%
Medium	25,000,000	28	33.9%
Small	2,500,000	110	32.8%
Very Small	0	55	1.9%

Figure 7. Illustrative Strata Boundaries for Union's C/I Projects

Stratum Size	Lower Threshold of Cumulative Gross Gas Savings (m^3)	Projects	Savings Represented (%)
Large	25,000,000	11	33.0%
Medium	5,000,000	49	34.6%
Small	1,500,000	195	27.9%
Very Small	0	214	4.5%

The "Very Small" projects—representing the bottom 1.9% of T1/R100 program savings and the bottom 4.5% of C/I program savings—are removed from the sample frame. These projects are small enough that the value of the information gained by evaluating them is not likely to be worth the cost. These projects should be adjusted by the Small Project stratum realization rate when re-introduced in the final sample analysis.

Step 3 estimates an appropriate variance for each stratum. Historical evaluation results indicate that CVs on project realization rates have been as low as 0.20 or as high as 0.40. However, typical CVs have been near 0.25. CVs are set at 0.30 for all strata in this example.

Step 4 allocates observations to each stratum. Figure 8 and Figure 9 indicate the sample sizes[64] and the assumptions used to allocate the samples when applying the calculations presented in Appendix B.

Figure 8. Illustrative Sample Allocation for Union's T1/R100 Projects

Stratum Size	Population Size	Sample Size	CV	T - value	FPC	Mean Gross Cumulative Gas Savings	Total Gross Cumulative Gas Savings	Stratum Weight
Large	10	7	0.3	1.94	0.58	88,950,000	889,500,000	0.32
Medium	28	7	0.3	1.94	0.88	34,339,286	961,500,000	0.35
Small	110	6	0.3	2.02	0.98	8,454,545	930,000,000	0.33
	148	20		1.73				1.00

[64] In previous program cycles when Union's custom programs were differentiated based on service contract rather than rate class, the differences between program sample sizes were much greater. Sample sizes will likely be more similar for the Union programs now that the programs differentiated based on rate class.

Figure 9. Illustrative Sample Allocation for Union's C/I Projects

Stratum Size	Population Size	Sample Size	CV	T - value	FPC	Mean Gross Cumulative Gas Savings	Total Gross Cumulative Gas Savings	Stratum Weight
Large	11	6	0.3	2.02	0.71	45,545,455	501,000,000	0.35
Medium	49	7	0.3	1.94	0.94	10,744,898	526,500,000	0.36
Small	195	7	0.3	1.94	0.98	2,176,923	424,500,000	0.29
	255	20		1.73				1.00

The sample allocations are restricted to less than 75% of the total population for the two Large Project strata. This restriction allows for some backup projects to exist for the Large Project strata so that if recruitment of the original sample is unsuccessful, backup projects can be used and the sample will likely not require re-stratification or re-allocation.

Step 5 determines criteria for assessing sample representativeness. Note that this is listed as an optional step; however, it can be important for ensuring that the most appropriate information is provided from this analysis for making regulatory decisions such as payment of incentives and future program decisions. While the sample methodology applies techniques to minimize the required sample sizes, the smaller samples are at an increased risk that a given random sample is not sufficiently representative for extrapolation to the population and used to assess whether savings targets have been met. This is why ensuring representativeness is an important step.

This example establishes simple criteria to ensure representativeness of the sample across market segment in the R1/T100 and the C/I program sample.[65] Several market segments are specified in the tracking database, and their proportions are shown in Figure 10 and Figure 11.

Figure 10. Illustrative Representativeness Analysis of Project Market Segment for Union'sT1/R100 Program

Project Market Segment	Large Projects Gross			Medium Projects Gross			Small Projects Gross		
	#	Cumulative m^3	%	#	Cumulative m^3	%	#	Cumulative m^3	%
Agriculture							6	54,000,000	6%
Food Services							1	12,000,000	1%
Healthcare							5	33,000,000	4%
Manufacturing	10	889,500,000	100%	27	919,500,000	96%	86	753,000,000	81%
Resource									
Utility				1	42,000,000	4%	12	78,000,000	8%
	10	889,500,000	100%	28	961,500,000	100%	110	930,000,000	100%

The main concern is that a randomly selected sample might under-represent the most important market segments, leading to a bias in program results. In these sample designs, less than ten

[65] Union and its sampling advisor may determine that no criteria are needed or that other criteria are needed based on judgment and assessment of actual program data.

sites may be drawn in a stratum; therefore, it is not impossible that this small sample size might be quite unrepresentative in some strata due to an unlucky sample draw. Increasing the sample sizes in each stratum could help resolve this issue, but the high cost of visiting each site and gathering the verification data makes this very expensive. As a result, this representativeness check should be considered.

In the T1/R100 program, manufacturing is clearly the dominant market segment and ensuring that a representative sample from this segment across size categories is all that may be needed; however, an evaluator may want to check to see if the random project selection (in the next step) provides some projects from non-manufacturing segments such as agriculture and utility market segments. The most significant risk is likely to occur in the small projects sample where manufacturing accounts for 78% of the projects and 81% of the savings. It could be possible to have an "extreme" sample occur in a random draw where non-manufacturing sites are "overly" represented.[66] The sample for this stratum is only six projects. If five of these projects are non-manufacturing when manufacturing accounts for 81% of the savings, this sample may not provide the information desired from this verification effort. A criteria that at least three of the projects in this stratum be manufacturing projects may represent the minimum needed to consider the sample representative overall.

Figure 11. Illustrative Representativeness Analysis of Project Market Segment for Union's C/I Program

Project Market	Large Projects Gross			Medium Projects Gross			Small Projects Gross		
Segment	#	Cumulative m^3	%	#	Cumulative m^3	%	#	Cumulative m^3	%
Agriculture				17	151,500,000	29%	56	121,500,000	29%
Education	2	144,000,000	29%	1	7,500,000	1%	13	36,000,000	8%
Entertainment							2	4,500,000	1%
Healthcare							19	33,000,000	8%
Manufacturing	9	357,000,000	71%	31	367,500,000	70%	99	214,500,000	51%
Multi-Family							2	4,500,000	1%
Resource							1	4,500,000	1%
Retail							1	1,500,000	0%
Transport							1	3,000,000	1%
Utility							1	1,500,000	0%
	11	501,000,000	100%	49	526,500,000	100%	195	424,500,000	100%

In the C/I program, the most important market segment is clearly manufacturing, followed by agriculture and education. To ensure that this is a representative sample, it may be important to be sure that the projects selected in the next step (random selection) contain some projects from each of these market segments. Manufacturing represents 65% of the overall savings. The agriculture and education market segments account for 19% and 13%, respectively, or 32% of total savings when taken together. Given a sample size of 20 overall, and no more than 7 in each stratum, a sample might be drawn that could be extreme and may not be an accurate

[66] What constitutes "overly" represented simply has to be defined by judgment exercised by the evaluator.

representation of the population. Again, the concern is the high cost of conducting the site visits, which argues against simply expanding the sample size or adding new strata. To ensure that manufacturing does not entirely dominate the sample, it might be good to set representativeness criteria, for example, that at least four sites be non-manufacturing sites.

Step 6 selects a random sample. The selection of the sample should be uniformly random within each stratum. This is accomplished by applying the RAND() function in Microsoft Excel and selecting the projects with the highest randomly assigned numbers to fulfill sample size requirements. The sample is reviewed to ensure that it meets any previously established criteria. Backup projects are also selected to replace any projects from the primary sample that are not successfully recruited.

Step 7 recruits the sample. Projects from the primary sample are only replaced after four recruitment attempts on four different dates. Projects that are not successfully recruited are documented before being replaced by backup projects.

These seven steps illustrate how the sample design methodology might be implemented using representative data. Following verification and evaluation of the sample, the sample data should be analyzed according to the realization rate methodology presented in Section 6 and according to the calculations presented in Appendix B.

5.6 Example Implementation of Sample Design Methodology (Enbridge)

This section demonstrates how the sample design methodology might be implemented for an example set of Enbridge program data. The data used for this example has been randomized and does not indicate historic program achievements that have undergone regulatory review in prior years. The data for this example is intended to be representative of a typical program year for illustrative purposes only. This example is for reference and does not preclude the judgment needed to understand and address the idiosyncrasies of actual program data.

This example applies the steps of the sample design process presented in Section 5.4.

Step 1 reviews the project tracking database for accuracy and quality. Of particular emphasis is a check on the processes used to produce the initial estimates for savings contained in the database and the contact information. This step is usually undertaken by the utility and is done to provide the third-party evaluator with the best information possible. As mentioned above, a more accurate tracking database will make it more likely that confidence and precision targets will be met. This example assumes that the tracking database has been reviewed.

Step 2 evaluates the population and defines strata. Gross *cumulative* gas savings measured in cubic meters (m^3) is the primary impact to be studied and should serve as the basis of the

sample design. Figure 12 and Figure 13 show representative project distributions of savings[67] for Enbridge's commercial and industrial programs, respectively. Analyzing the distribution of project sizes indicates that size-based stratification should produce sampling efficiencies. Other categorical bases for stratification are not chosen for this example.

Figure 12. Illustrative Distribution of Savings for Enbridge Commercial Projects

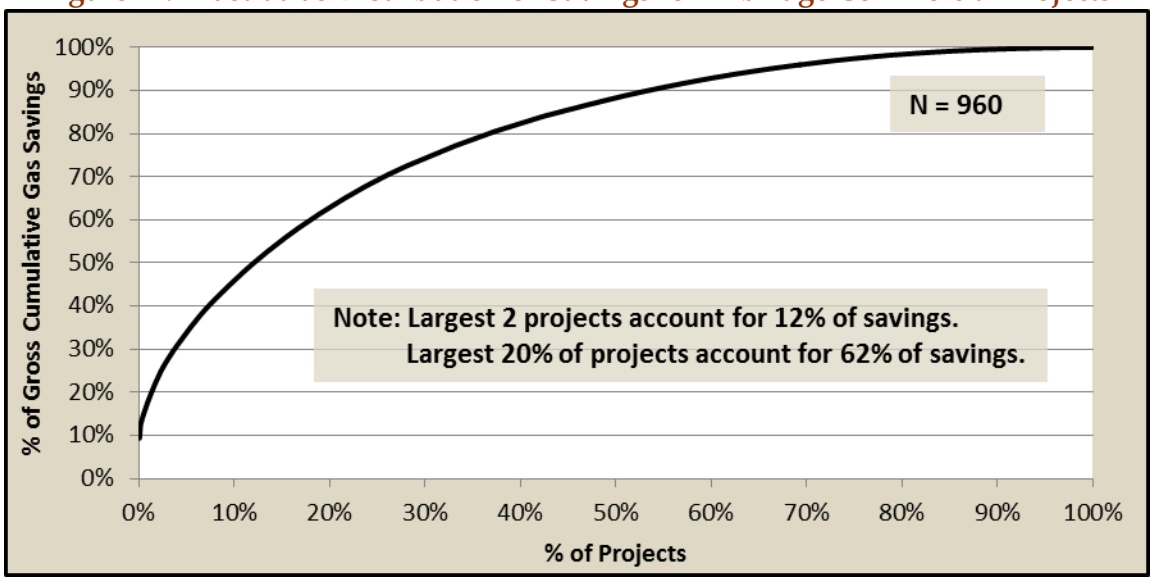
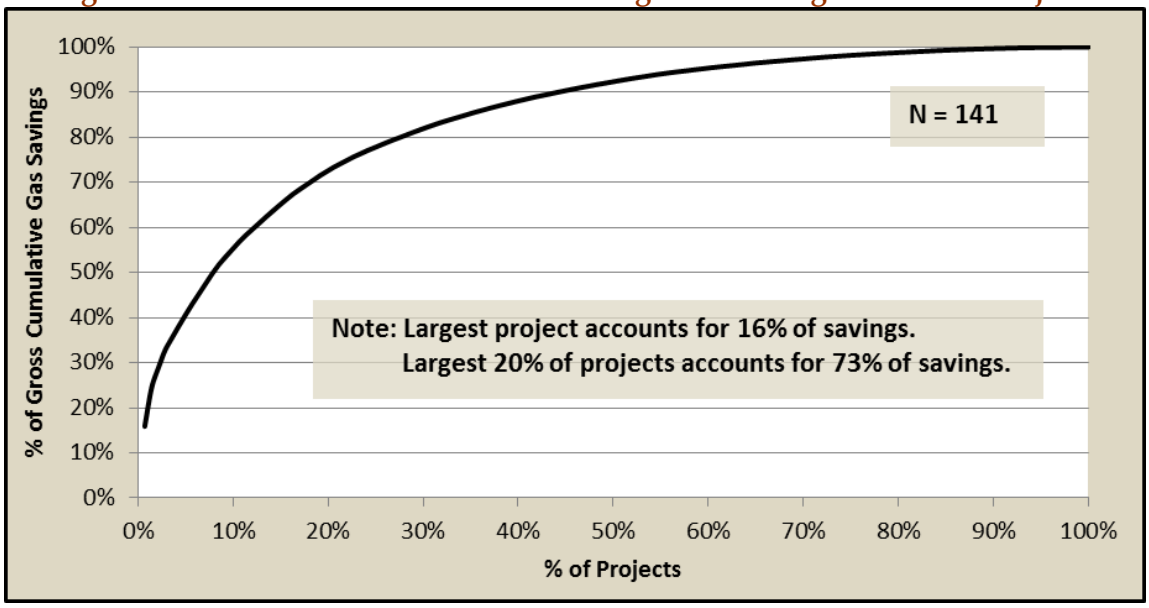


Figure 13. Illustrative Distribution of Savings for Enbridge Industrial Projects



[67] The initial manual produced in November, 2012 used net gas savings in the examples. In this revised report, the example analyses are performed on cumulative gross savings values to correctly illustrate how that the sampling and the application of population-wide realization rates for the utilities should be performed in these sampling analyses.

The sensitivity to sample sizes is investigated to determine appropriate savings thresholds for strata bounds. Since the commercial program has a relatively large number of projects, it is necessary to balance the effects of strata weight with the effects of finite population correction when determining the threshold for the Large Project stratum. Figure 14 and Figure 15 show illustrative strata boundaries for Enbridge's commercial and industrial programs, respectively.

Figure 14. Illustrative Strata Boundaries for Enbridge Commercial Projects

Stratum Size	Lower Threshold of Cumulative Gross Gas Savings (m^3)	Projects	Savings Represented (%)
Large	8,000,000	9	17.6%
Medium	2,000,000	153	40.7%
Small	400,000	479	36.9%
Very Small	0	319	4.8%

Figure 15. Illustrative Strata Boundaries for Enbridge Industrial Projects

Stratum Size	Lower Threshold of Cumulative Gross Gas Savings (m^3)	Projects	Savings Represented (%)
Large	14,000,000	8	40.5%
Medium	5,000,000	22	32.8%
Small	500,000	79	25.1%
Very Small	0	32	1.5%

The "Very Small" projects—representing the bottom 4.8% of commercial program savings and the bottom 1.5% of industrial program savings—are removed from the sample frame. These projects are small enough that the value of the information gained by evaluating them is not likely to be worth the cost. These projects should be adjusted by the Small Project stratum realization rate when re-introduced in the final sample analysis.

Step 3 estimates an appropriate variance for each stratum. Historical evaluation results indicate that CVs on project realization rates have been very low, sometimes less than 0.10. However, applying CVs less than 0.30 is not recommended in order to ensure sample sizes sufficient for robust results and to allow for increasing variances that may result from evolving measurement approaches and program participation. CVs are set at 0.30 for all strata in this example.

Step 4 allocates observations to each stratum. Figure 16 and Figure 17 indicate the sample sizes and the assumptions used to allocate the samples when applying the calculations presented in Appendix B.

Figure 16. Illustrative Sample Allocation for Enbridge's Commercial Program

Stratum Size	Population Size	Sample Size	CV	T - value	FPC	Mean Gross Cumulative Gas Savings	Total Gross Cumulative Gas Savings	Stratum Weight
Large	9	5	0.3	2.13	0.71	751,111	6,760,000	0.18
Medium	98	8	0.3	1.89	0.97	110,384	13,798,000	0.37
Small	590	11	0.3	1.81	0.99	29,766	16,758,000	0.45
	697	24		1.71				1.00

Figure 17. Illustrative Sample Allocation for Enbridge's Industrial Program

Stratum Size	Population Size	Sample Size	CV	T - value	FPC	Mean Gross Cumulative Gas Savings	Total Gross Cumulative Gas Savings	Stratum Weight
Large	8	6	0.3	2.02	0.41	33,321,429	233,250,000	0.41
Medium	22	6	0.3	2.13	0.87	8,590,909	189,000,000	0.33
Small	79	5	0.3	2.35	0.97	1,809,938	144,795,000	0.26
	109	17		1.75				1.00

The key reason that the required sample size is smaller for the industrial program than the commercial program is that a larger fraction of the savings is concentrated in a smaller number of projects for the industrial program. The sample allocations are restricted to less than 75% of the total population for the two Large Project strata. This restriction allows for some backup projects to exist for the Large Project strata so that if recruitment of the original sample is unsuccessful, backup projects can be used and the sample will likely not require re-stratification or re-allocation.

Step 5 determines criteria for assessing sample representativeness. Note that this is listed as an optional step ; however, it can be important for ensuring that the most appropriate information is provided from this analysis for making regulatory decisions such as payment of incentives and future program decisions. While the sample methodology applies techniques to minimize the required sample sizes, the smaller samples are at an increased risk that a given random sample is not sufficiently representative for extrapolation to the population and used to assess whether savings targets have been met. This is why ensuring representativeness is an important step.

This example establishes a simple criterion to ensure representativeness of load type in the commercial program sample.[68] Three load types are specified in the tracking database, and their proportions are shown in Figure 18.

[68] Enbridge and its sampling advisor may determine that no criteria are needed or that other criteria are needed based on judgment and assessment of actual program data.

Figure 18. Illustrative Analysis of Project Load Types for Enbridge's Commercial Program

Project Market Segment	Large Projects #	Large Projects Gross Cumulative m3	Large Projects %	Medium Projects #	Medium Projects Gross Cumulative m3	Medium Projects %	Small Projects #	Small Projects Gross Cumulative m3	Small Projects %
Space Heating	7	202,200,000	92%	135	438,300,000	86%	416	414,660,000	89%
Water Heating	1	10,500,000	5%	5	16,500,000	3%	53	37,440,000	8%
Combined	1	8,100,000	4%	13	55,800,000	11%	10	11,670,000	3%
Grand Total	9	220,800,000	100%	153	510,600,000	100%	479	463,770,000	100%

The main concern is that a randomly selected sample might over-represent water heating to the detriment of properly representing space heating projects simply due to an unlucky draw of insufficiently representative projects. As example criteria, it might be reasonable to require that space heating projects must account for at least 70% of the savings in each stratum. A sample that does not meet these criteria would be viewed as unrepresentative and would be discarded and re-selected.

Step 6 selects a random sample. The selection of the sample should be uniformly random within each stratum. This is accomplished by applying the RAND() function in Microsoft Excel and selecting the projects with the highest randomly assigned numbers to fulfill sample size requirements. The sample is reviewed to ensure that it meets any previously established criteria. Backup projects are also selected to replace any projects from the primary sample that are not successfully recruited.

Step 7 recruits the sample. Projects from the primary sample are only replaced after four recruitment attempts on four different dates. Projects that are not successfully recruited are documented before being replaced by backup projects.

These seven steps illustrate how the sample design methodology might be implemented using representative data. Following verification and evaluation of the sample, the sample data should be analyzed according to the realization rate methodology presented in Section 6 and according to the calculations presented in Appendix B.

5.7 Summary of Sample Design Methodology

The sample design methodology described in this section is meant to apply advanced industry practices to create a cost-efficient sample by leveraging preexisting project and program information to the greatest extent possible. The methodology can be described as employing a "stratified ratio-estimation" approach. The sample is administered in two stages to make the best use of early observations that can be collected prior to completion of the program year. The methodology provides a step-by-step description of sample design tasks, but leaves flexibility to accommodate program changes in future years and cycles.

6. Recommended Realization Rate Methodology

This section describes the recommended methodology for determining realization rates and achieved confidence and precision based on sample observations of custom DSM programs for Union and Enbridge. Section 6.1 describes the approach to determine verified realization rates. Section 6.2 describes the approach to determine the precision on the realization rate and total savings achieved by the sample. Section 6.3 discusses several potential adjustments that may be needed to ensure that the results appropriately characterize the population and provide the information needed by the utilities and stakeholders.

It is important ensure the quality of sample observation data prior to calculating achieved realization rates and savings. Data quality issues can sometimes be discovered when analyzing the sample, but it can be costly to correct the data at that point. Undetected data quality issues would result in inaccuracies of total savings and precision estimates.

6.1 Determining Verified Realization Rates

Gross realization rates should be calculated for each stratum sample and applied to each respective stratum population when estimating total gross cumulative gas savings.[69]

Applying gross realization rates to population strata is more complicated than assessing the results in a simple random sample without strata, but it is necessary when efficiencies are sought through stratification.[70]Again, efficiencies are important in this application due to the high cost of gathering the verification data at each sample site. Lohr notes:

> *The population total is the [sum across all strata of the estimated stratum population mean times the stratum population size]…This is a weighted average of the sample stratum averages; the weights are the relative sizes of the strata. To use stratified sampling, the sizes or relative sizes of the strata must be known.* [71]

Also, Wadsworth notes:

> *The estimator of the total of a stratified population can be expressed as the sum of strata of estimators of the individual stratum totals. This representation suggests the valid generalization that the estimator of the total in a stratum need not be limited to the expansion estimator, but could be any appropriate estimator of the population in the stratum, including a ratio*

[69] Ultimately, adjusted gross savings can be converted to adjusted net savings (i.e. by applying a program net-to-gross ratio to the adjusted program gross savings). However, that would occur outside of (i.e. after) the application of the sampling work discussed in this report.

[70] There are examples in the evaluation literature where strata weights have not been used in the calculation of the mean realization weight. This is clearly an oversight in these evaluations as it is a simple matter to weight the mean ratios of each stratum by the appropriate stratum weight (i.e., the proportion of the population in that stratum).

[71] Lohr, S. L., "Sampling: Design and Analysis," Second Edition. 2010, p. 69.

estimator…then an estimate of the total in a stratified population may be constructed as a sum over strata. [72]

These are standard procedures for developing population estimates from a stratified sample. The methods for estimating the population parameters must take into account the strata weights when stratification is used. The calculations needed to develop a verified gross realization rate from stratified sample data are shown in Appendix B. This approach is based on widely recognized methods published by Lohr.[73]

This approach for determining gross realization rates is consistent with the recommended sample design methodology presented in Section 5.

6.2 *Determining Achieved Confidence & Precision*

A precision level cannot be calculated without first establishing the confidence level. The calculation for both confidence and precision comes from the same basic equation. Either confidence or precision is first established, then the other is solved for. For example, a precision of +/- 10% implies that the stated confidence level should span +/- 10% from the mean estimate. The confidence may turn out to be 90%, 82% or another value. The confidence level is more typically established and the precision is solved for. For example, the level of precision achieved at a 90% level of confidence can be calculated and may turn out to be 10%, 12%, 15% or some other number (as illustrated in Appendix A). Regardless, the calculating confidence and precision are part of the same equation and one cannot be estimated without establishing the other. Misunderstanding this basic concept frequently leads to problems in presenting and discussing evaluation results in the industry. Additional discussion on confidence and precision can be found in Appendix A.

Confidence and precision calculations also have to take into account the fact that a stratified random sample has been used. The equations for calculating confidence and precision from a stratified sample design are shown in Appendix B. This approach for determining confidence and precision is consistent with the recommended sampling methodology in Section 5, and it is consistent with the population realization rate and savings estimates described in Section 6.1.

Communications with the TEC indicated that they were interested in both the likelihood that savings exceeds a given value and the likelihood that it falls above a given value. As a result, the recommendation is to report achieved confidence and precision in three ways:[74]

1. Achieved precision corresponding to 90% one-sided confidence on the lower bound
2. Achieved precision corresponding to 90% one-sided confidence on the upper bound[75]

[72] Wadsworth, H.M., "Handbook of Statistical Methods for Engineers and Scientists," 1990, p. 9.25.

[73] Lohr, S. L., "Sampling: Design and Analysis," Second Edition.2010. (Sections 4.1-4.5)

[74] The achieved precision is a result of analyzing the sample data, and will usually differ to some extent from the targeted precision applied in designing the sample.

3. Achieved precision corresponding to a 90% two-sided confidence interval

Appendix A provides additional explanation and illustrative examples for the reporting of confidence and precision in the estimated realization rate. The Figures in Appendix A are intended to clarify the interpretation of confidence and precision in making decisions based on the estimated realization rate.

6.3 Sample Adjustments & Related Issues

This section discusses several sampling adjustments that may be needed to accurately synthesize the total population realization rate and savings estimates. The following three types of adjustments are discussed:

1. Treatment of outliers and influential observations
2. Replacing sample projects
3. Post-stratification

Appropriately treating outliers and influential observations is important in accurately estimating the realized savings for DSM programs. Parties to a discussion of estimating program savings should understand appropriate treatment of outliers and influential observations when estimates are based on a sample of the population.

Treatment of Outliers & Influential Observations

This section first presents a conceptual discussion. Following this discussion, an example from a recent Union custom program evaluation is presented. Most statistical analyses should examine the data for outliers and test to determine whether these outliers may be "influential observations" that can skew the accuracy of a sample. Kennedy states the rationale for treating outliers:

> *The rationale for looking for outliers is that they may have a strong influence on the estimates…an influence that may not be desired.* [76]

In other words, the reason for looking for evaluating outliers is that there may be a sample case drawn that is well outside the expected bounds of the distribution and that this observation may exert undue influence on the estimates of the analysis (i.e., an influential observation). Osborne and Overbay further describe the effect of outliers:

> *The presence of outliers can lead to inflated error rates and substantial distortions of parameter and statistic estimates when using either parametric or nonparametric tests (e.g., Zimmerman,*

[75] Achieved precision of the upper bound represents a simple inversion of the confidence interval for the lower bound. Reporting on the upper bound is intended to facilitate an understanding that sampling uncertainties can just as likely lead to underestimation of the realization rate and therefore underestimating overall program savings as they are to result in overestimates.

[76] Kennedy, P. "A Guide to Econometrics." Third Edition. MIT Press, 1992, p. 279.

> *1994, 1995, 1998). Casual observation of the literature suggests that researchers rarely report checking for outliers of any sort.* [77]

The issue is whether it is appropriate for a single observation to swing the overall results in a substantial manner.[78] If such an observation is found, then further study is needed to determine the most appropriate course of action. In general, a sample of 10 from a population of 100 projects implies that each sample point represents 10 projects. However, if a selected sample point is truly a unique case and does not represent other projects in the population, then an adjustment may be warranted. Osborne and Overbay go on to state:

> *[The appropriate treatment] depends in large part on why an outlier is in the data in the first place. Where outliers are illegitimately included in the data, it is only common sense that those data points should be removed… Few should disagree with that statement.*

The sample analysis should seek to determine whether or not outliers and influential observations can be viewed as representative members of the main population upon which population estimates may be inferred. Barnett and Lewis note:[79]

> *If they are not [suitable]…they may frustrate attempts to draw inferences about the original (main) population.*

One example can be taken from the analysis of the sample observation in Union's 2011 custom program. Two outliers were identified in the Distribution Contract (DC) custom program. One verified project observed a gas savings realization rate of 3.75 and a second project observed a realization rate of 0.18. A sensitivity analysis tested for the influence of these two observations by removing[80] them and noting the changes in results.[81]

The estimated overall realization rate for gas savings when including both observations was 1.25. This is a relatively high realization rate when compared to evaluation efforts across North America, but not an unheard of result. Excluding the high observation lowered the estimated overall estimate from 1.25 to 1.05. Excluding the low observation raised the overall estimate

---

[77] Osborne, J., Overbay, A. "The Power of Outliers and Why Researchers Should Always Check for Them."2004 Practical Assessment, Research & Evaluation, volume 9, section 6. Link: http://pareonline.net/getvn.asp?v=9&n=6

[78] A simple intuitive example of the impacts an outlier can have on a statistical analysis can be found in a Wikipedia contribution (8/20/2012): *Naive interpretation of statistics derived from data sets that include outliers may be misleading. For example, if one is calculating the average temperature of 10 objects in a room, and nine of them are between 20 and 25 °C, but an oven is at 175 °C, the median of the data could be between 20 and 25 °C but the mean temperature will be between 35.5 and 40 °C. In this case, the median better reflects the temperature of a randomly sampled object than the mean; however, naively interpreting the mean as "a typical sample", equivalent to the median, is incorrect. As illustrated in this case, outliers may be indicative of data points that belong to a different population than the rest of the sample set.*

[79] Barnett, V., Lewis, T., "Outliers in Statistical Data." Wiley Series in Probability & Statistics, 1998/1994.

[80] Removing or excluding an outlier entails isolating the sample point in a unique stratum such that the sample point still counts in the analysis, but it is not used for extrapolating results for the un-sampled population.

[81] Note that some observations may be identified as outliers but do not significantly influence the analysis results.

from 1.25 to 1.32. Excluding both outliers produced an overall realization rate on gas savings of 1.11.

Discussions were held with Union concerning the two outlier observations. It is important not to exclude an observation without examining the reasons that may contribute to the observation's extreme value. If the observation is representative of other projects in the population, it should be left in. If it can be shown to result from a one-time construct and is not likely to be replicated by other members of the population, then exclusion of this observation should be considered. The discussions with Union indicated that both observations were likely due to unique calculation issues and technologies involved.

The most conservative position in treating this outlier issue was taken—the high observation was removed and the low observation was retained in the sample data set. This produces the lowest overall program realization rate given the choices in addressing the identified outliers. However, removing outliers in strata with small sample sizes may also adversely affect the confidence and precision results and the sample may require augmentation to achieve confidence and precision targets.

Projects that implement new technologies—whose savings estimates have had less validation—or certain technology classes that are complex and difficult to estimate for the tracking database may be at an increased likelihood to result in outlier realization rates. Identifying such projects in the program tracking database could help isolate them and reduce their chance of skewing program estimates. These projects could be placed into a separate category with different confidence and precision targets for new technologies. Any projects that are truly unique should be identified and addressed during sample design. These steps would not eliminate these projects in terms of their contribution to overall program savings, but would allow for appropriate methods to more accurately estimate program savings. If sampled, these unique projects should not be considered representative of other projects in the main program. As a result, addressing this issue in advance could improve the sample analysis and the resulting program estimates.

Replacing Sample Projects

The final recruited sample should be analyzed and summarized, especially when replacement projects are substituted into the originally selected sample. Recruiters should document the reasons for unsuccessful recruitment of original sample members. Replacement samples should always be selected in priority based on the assigned random number, and full effort should be made to recruit selected replacements before substituting other replacements. If recruitment rates are very poor, this may introduce a significant non-response bias. Low recruitment rates should be investigated and documented, and recommendations may be made to improve recruitment in subsequent evaluation years.

Post-Stratification

If a sample did not achieve the desired confidence and precision and the stratification basis is thought to be sub-optimal, post-stratification may be used to retrospectively re-stratify a sample along more appropriate dimensions to demonstrate an improved precision achieved by the sample. Often, post-stratification will not improve achieved precision, especially at relatively small sample sizes; however, under certain circumstances this technique may be useful. The Ontario Power Authority notes that:

> *A technique known as post-stratification may be used to develop estimates about sub-populations after the study is complete and can be used if characteristics about the sub-populations are unknown at the time the study in conducted.*
>
> *This advanced technique should be reserved for special situations and utilized only after careful consideration of other options and well documented in the experimental approach of the Draft Evaluation Plan.* [82]

Post-stratification should not be used on a normal basis, and if necessary should inform subsequent program evaluation cycles to improve the sample frame and prevent the need for post-stratification in future years.

6.4 Summary of Realization Rate Methodology

This section presents the method for calculating verified ex-post realization rates as well as for appropriately calculating the confidence and precision levels for the estimated realization rate and overall program savings. It also discusses three issues that can lead to adjustments to the sample and recalculation of the realization rate along with confidence and precision levels. There are several important concepts presented in this section:

- The program realization rate is inferred from the sample observations based on the separate realization rates for each stratum.
- The realization rate calculations should apply the strata weights to accurately interpret sample observations. This adds a bit of complexity, but no alternate application of the observed data would be appropriate. This is considered standard practice in the application of a stratification approach in statistics.
- There are some important and legitimate considerations that should be examined when inferring estimates for a population from an observed sample. The following three factors are discussed in this section:
    1. Outliers and influential observations
    2. Replacement projects when data cannot be gathered from the originally sampled project

[82] "EM&V Protocols and Requirements: 2011-2014." Ontario Power Authority. March 2011, p. 130.

3. Post-stratification to provide higher precision and greater confidence in the results

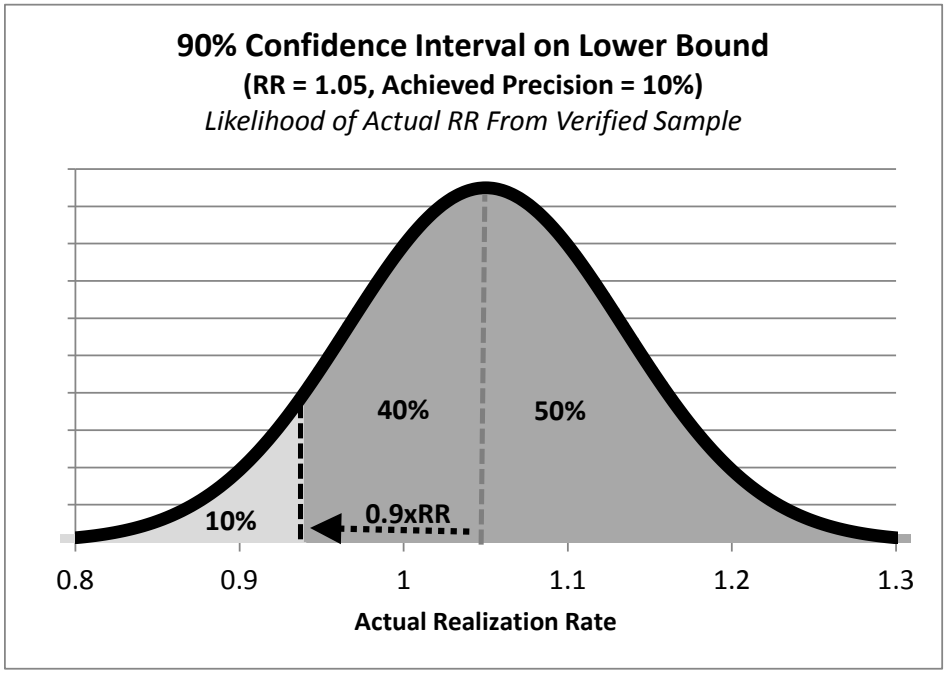
The equations needed to calculate the realization rates and achieve confidence and precision from the sample data are contained in Appendix B.

Appendix A. Explanatory Note on Confidence & Precision

The level of certainty associated with a statistical sample is most often stated in terms of a confidence interval. A confidence interval contains two components: confidence level and precision. Confidence level indicates the likelihood that an actual variable either exceeds a value (i.e., one-sided confidence) or falls within a range (i.e., two-sided confidence). Precision[83] indicates the bounding values of the corresponding confidence level. Confidence and precision are both necessary to sufficiently describe a confidence interval.[84]

At the time of this report, the target confidence interval for the design of the sample is established as 90/10 one-sided.[85] Figure 19 illustrates a 90% one-sided confidence interval with 10% precision for a sample whose realization rate (RR) is estimated to be 1.05.

Figure 19. Illustration of a 90% One-Sided Confidence Interval on the Lower Bound



[83] Relative precision (e.g., 10% of the estimate) is most often used to set the precision as a percentage of the estimated value rather than in absolute terms.

[84] Also, the shape (i.e., one-sided or two-sided) is often used to fully specify the confidence interval.

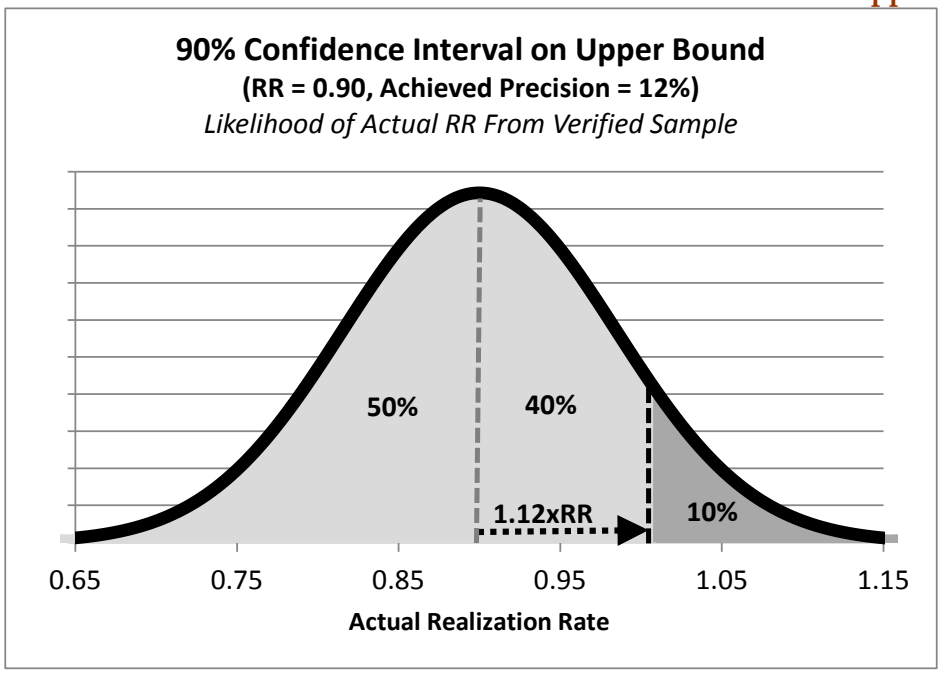
[85] Based on October 25, 2012 Technical Evaluation Committee decision the sample design should be based on a 90/10 one-sided confidence interval. Reporting of achieved confidence and precision should present the precision achieved for both the 90% one-sided and 90% two-sided intervals.

Reading off of Figure 19, this confidence interval can be interpreted as showing that:[86]

- There is a 10% likelihood that the actual value is less than 10% below the mean sample estimate of 1.05.
- There is a 40% likelihood that the actual value falls between 10% below the sample estimate and the sample estimate of 1.05.
- There is a 50% likelihood that the actual value exceeds the sample estimate of 1.05.

The reporting recommendations in Section 6.2 of the main report also call for the reporting of a one-tailed test around an upper bound and a two-tailed test at a 90% confidence level. These are illustrated in Figure 20 and Figure 21. Figure 20 illustrates a 90% one-sided confidence interval on the upper bound. For this illustration a different realization rate estimate is use that was used in Figure 19. In this case, the estimated realization rate is 0.90 and the level of precision achieved at the 90% confidence level is observed from the sample to be 12%. This confidence interval illustrates that the actual value has a 10% likelihood of exceeding the estimated realization rate of 0.90 plus 12% (i.e., exceeding a realization rate 1.01). This likelihood is illustrated by the dark shaded portion of the distribution in the Figure.

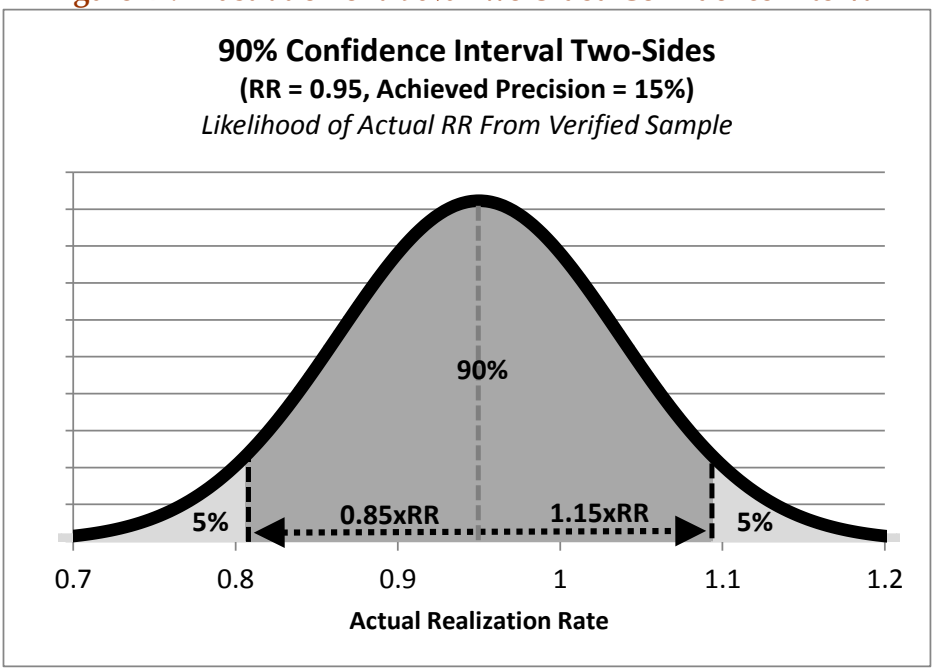
Figure 20. Illustration of a 90% One-Sided Confidence Interval on the Upper Bound



[86] This interpretation of the confidence interval is based on statistical inference, which assumes that the sample provides an adequate representation of the population.

Figure 21 illustrates a 90% two-sided confidence interval on a sample whose realization rate is observed to be 0.95 and whose achieved precision is 15%. The dark shaded area in the middle of the distribution represents the 90% confidence level that the actual value would fall between the bounds set plus or minus 15% of the observed sample estimate. There is only a 5% likelihood that the actual value would fall below the lower bound.

Figure 21. Illustration of a 90% Two-Sided Confidence Interval



Appendix B presents the detailed calculation methods for determining the confidence and precision achieved by a sample.

Appendix B. Calculation Methods & Equations

B.1 Calculating Target Sample Confidence & Precision from Assumed CV

(Note: The formulae in this appendix are based on application of Lohr[87] and Cochran,[88] and are adapted to the vocabulary of the stratified realization rate problem of efficiency program evaluation.)

The standard error of the total savings of stratum h based on tracked ex ante savings[89] is given by,

$$SE'_h = FPC_h \times \frac{CV_h}{\sqrt{n_h}} \times TS'_h$$

Where CV_h[90] is the estimated coefficient of variation in stratum h, defined as the expected stratum standard deviation divided by the expected stratum mean.[91] Where FPC_h is the finite population correction factor of stratum h, n_h is the sample size of stratum h, and TS'_h is the tracked ex ante total savings in stratum h.[92] FPC_h is given by,

$$FPC_h = \sqrt{\frac{N_h - n_h}{N_h - 1}}$$

Where N_h is the population size of stratum h. The relative precision at the stated confidence level of stratum h is given by,

$$RP'_h = t_h \times \frac{SE'_h}{TS'_h} \times 100\%$$

Where t_h is the t-value derived from the confidence requirement and the sample size of stratum h. The overall standard error can be calculated by aggregating the sample according to each stratum's weighting (i.e., expected percent contribution to total program savings). The overall standard error of the tracked ex ante total savings of the program is given by,

[87] Lohr, S. L., "Sampling: Design and Analysis," Second Edition, 2010.
[88] Cochran, W. G., "Sampling Techniques," Third Edition, 1977.
[89] The prime symbol (apostrophe) is used to indicate that these values are based on tracked ex ante values rather than verified ex post values.
[90] In cases of ratio estimation, the error ratio is substituted for the coefficient of variation.
[91] The coefficient of variation may be based on savings or realization rate, as in the case of ratio estimation.
[92] Total tracked ex ante is not necessarily required to compute relative precision since this term is also in the denominator of the relative precision calculation.

$$SE'_P = \sqrt{\sum_h {SE_h}^2}$$

The overall relative precision at the stated confidence level is given by,

$$RP'_P = t_P \times \frac{SE'_P}{TS'_P} \times 100\%$$

Where t_P is the t-value derived from the confidence requirement and the overall sample size in the population, and TS'_P is the estimated total savings across all strata based on verified ex post savings.

B.2 Calculating Achieved Realization Rates

Defining $x_{i,h}$ as the tracked ex ante estimate and $y_{i,h}$ as the verified ex post estimate of a single sample point i in stratum h, the effective realization rate of a single sample point i in stratum h is given by,

$$RR_{i,h} = \frac{y_{i,h}}{x_{i,h}}$$

The stratum sample realization rate of stratum h is the sum of all verified ex post savings in the sample of stratum h divided by the sum of all tracked ex ante savings in the sample of stratum h, given by,

$$RR_h = \frac{\sum_{i \in h} y_{i,h}}{\sum_{i \in h} x_{i,h}}$$

In stratified ratio estimation, the stratum realization rate should be applied to the tracked ex ante estimates of each member j[93] of the full population of stratum h to produce the total savings estimate for stratum h. The verified total savings estimate for stratum h is the sum of all tracked ex ante estimates in stratum h multiplied by the stratum realization rate, given by,

$$TS_h = RR_h \times \sum_{j \in h} x_{j,h}$$

[93] Note that i members of the sample are a subset of j total members of the applicable population.

The verified total savings of the program can be calculated by aggregating strata results. The program verified total savings estimate is given by,

$$TS_P = \sum_h TS_h$$

The overall realization rate across all strata is the verified total savings of the program divided by the tracked ex ante total savings of the program, given by,

$$RR_P = \frac{TS_P}{TS'_P}$$

B.3 Calculating Achieved Sample Confidence & Precision

A predicted estimate can be made for each member of stratum h based on the stratum realization rate, where the predicted estimate is the tracked ex ante estimate of each member of the stratum multiplied by the stratum realization rate. A residual error can be calculated for each sample point in stratum h based on the difference between the verified ex post savings of the sample point and the predicted estimate. The residual of each sampled point is given by,

$$e_{i,h} = y_{i,h} - RR_h \times x_{i,h}$$

The sample variance[94] of the verified total savings in stratum h is derived from the stratum residuals, given by:

$$V_h = \frac{1}{n_h - 1} \sum_{i \in h} e_{i,h}^2$$

The standard error of the sample of stratum h can be calculated using the stratum sample variance and the finite population correction factor. The standard error of the verified total savings of stratum h is given by,

$$SE_h = FPC_h \times \frac{\sqrt{V_h}}{\sqrt{n_h}} \times N_h$$

[94] Sample variance is based on residuals of the verified measurement compared to the predicted estimate using the stratum realization rate when applying ratio estimation.

The relative precision for the stated confidence level of the verified estimate of stratum h is given by,

$$RP_h = t_h \times \frac{SE_h}{TS_h} \times 100\%$$

The resulting confidence interval can be stated in terms of the realization rate or the total estimate. The absolute two-sided confidence interval for the stratum realization rate and verified total savings of stratum h is given by,

$$RR_h \pm (RR_h \times RP_h) \quad and \quad TS_h \pm (TS_h \times RP_h)$$

The absolute one-sided confidence interval for the stratum realization rate and verified total savings of stratum h is given by,

$$> RR_h - (RR_h \times RP_h) \quad and \quad > TS_h - (TS_h \times RP_h)$$

The standard error of the verified total savings of the program is given by,

$$SE_P = \sqrt{\sum_h SE_h^2}$$

The overall relative precision at the stated confidence level is given by,

$$RP_P = t_P \times \frac{SE_P}{TS_P} \times 100\%$$

The absolute two-sided confidence interval for the overall program realization rate and verified total savings of the program is given by,

$$RR_P \pm (RR_P \times RP_P) \quad and \quad TS_P \pm (TS_P \times RP_P)$$

The absolute one-sided confidence interval for the overall program realization rate and verified total savings of the program is given by,

$$> RR_P - (RR_P \times RP_P) \quad and \quad > TS_P - (TS_P \times RP_P)$$

Appendix C. Summaries of Custom C&I Samples in Selected Jurisdictions

This appendix presents brief summaries of the sampling approaches used in custom commercial and industrial (C&I) programs in selected jurisdictions. The reviewed approaches are all contained within publicly available documents. Because the reviewed documents contain varying degrees of detail and explanation, the Navigant team applied its best interpretation of these documents to synthesize the available information in a consistent manner. Eight jurisdictions are discussed below. Published information on the sampling procedures allowed for a useful summary to be produced.

C.1 Summary from Illinois (ComEd)

The Commonwealth Edison Company (ComEd) Smart Ideas for Your Business program offers all eligible commercial and industrial customers financial incentives for upgrading their facilities with energy-efficient equipment. The program offers prescriptive incentives, available for qualified equipment commonly installed as part of retrofit and equipment replacement projects, or custom incentives, available for less common and more complex energy-saving measures. Examples of custom projects include heating, ventilating, and air conditioning (HVAC) measures (such as chiller upgrades and centralized thermostat control systems), large commercial refrigeration measures, air compressor system upgrades, high-rise building domestic water pumping systems, industrial process renovations, and non-prescriptive lighting measures. In 2011, the custom incentive levels were $0.03/kilowatt-hour (kWh) for equipment with less than a five-year life and $0.07/kWh for equipment with a five-year life or greater.[95] These incentive levels were applied for the first $100,000 in incentives and then reduced by half for the next $100,000, up to the project cost cap. In 2011, ComEd provided financial incentives to 887 projects. Of these, 32 projects were selected for evaluation to achieve confidence and precision targets of 90% and 8% over the three-year program.[96]

A two-stage sampling methodology was implemented, with the first projects being sampled in April of 2011 and the remaining projects sampled in July. The sampling approach stratified the population of projects by project size. All custom projects were sorted into three strata based on *ex ante* energy (kWh) savings, such that each stratum contained one-third of the total claimed energy savings.[97] The evaluation sample was drawn to represent the population distribution by stratum. Figure 22 shows the total number of projects and the evaluation sample by stratum. This sample represents 100% of the population's claimed energy savings in the first stratum,

[95] Any project involving Energy Management System programming is eligible for the $0.03/kWh incentive. To receive the $0.07/kWh custom incentive, equipment must have a minimum payback of one year and a maximum payback of seven years.

[96] A thirty-third project had been selected but after the site-visit it was moved into the following program year (PY4).

[97] Note that ComEd's custom program application does not require that applicants submit an estimate of savings, suggesting that the claimed savings may be underestimated. In addition, more projects may be assigned to stratum 3, resulting in a less precise estimation of *ex post* gross impacts.

59% in the second, and 5% in the third. In total, the 32 projects represent 45% of the program's custom projects' *ex ante* energy savings.

Figure 22. ComEd 2011 C&I Sample Summary

Sampling Stratum	Total Number of Projects	Evaluation Sample
1	2	2
2	27	15
3	858	15
Total	**887**	**32**

Source: Navigant Review of Evaluation Report[98]

C.2 *Summary from Michigan (DTE Energy)*

The DTE Energy C&I non-prescriptive program offers business customers financial incentives for the installation of "innovative and unique" energy efficiency equipment and controls. Examples of custom measures include energy management system controls, variable-speed air compressors, and ultrasonic HVAC humidification systems. Ineligible customer measures include on-site electricity generation, renewable energy, peak-shifting, fuel switching, or changes in operational/maintenance practices that do not involve capital costs. The custom incentive levels are $0.08/kWh, based on the first year of estimated energy savings, up to 50% of the project cost. Projects require a one-year minimum payback and an eight- year maximum payback.

In 2010, DTE Energy provided financial incentives for 515 energy efficiency measures associated with 381 unique projects. Of these projects, 56 were selected for evaluation to achieve confidence and precision targets of 90% and 10%, respectively, at the program level. This sample of 56 was based on a proportional sampling of measures from each of the three major technology groups: custom lighting, custom electric and custom gas.[99] Figure 23 shows the number of energy efficiency measures, unique projects, and evaluation sample size by group. The sample of custom lighting measures, custom electric measures, and custom gas measures represents 60%, 45%, and 90% of *ex ante* gross energy savings, respectively, for the population.

[98] "Evaluation Report: Smart Ideas for Your Business Custom Program." (Program Cycle 2010-2011.) Commonwealth Edison Company. Prepared by Navigant Consulting, Incorporated. May 16, 2012.

[99] Due to the small sample of "custom electric", several additional measure types were consolidated into this group to avoid a potential distortion in the realization rate. For example, custom HVAC, custom motors, and measures installed through a grocery RFP are included in the "custom electric" category.

Figure 23. DTE Energy 2010 Custom C&I Sample Summary

Sampling Stratum	Total Number of Measures	Total Number of Projects	Evaluation Sample
Custom Lighting	321	252	27
Custom Electric	150	93	9
Custom Gas	44	36	20
Total	**515**	**381**	**56**

Source: Navigant Review of Evaluation Report[100]

C.3 Summary from Massachusetts (National Grid, NSTAR, and Western Massachusetts Electric Company)

The C&I energy efficiency program run by the Massachusetts Program Administrators offers financial incentives to business customers for installing energy-efficient equipment. Custom projects are categorized as either a comprehensive design (CD) project or a comprehensive chiller (CC) project. CD projects typically involve the new construction of commercial, industrial, or municipal buildings that include at least four energy conservation measures (ECMs) that achieve a minimum of 20% energy savings relative to code.[101] CC projects typically involve the installation of a new chiller and multiple other ECMs in an existing building that achieve a minimum of 20% savings.

In 2008 and 2009, 25 custom projects were installed in National Grid, NSTAR, and Western Massachusetts Electric Company (WMECO) service territories.[102] Custom projects were stratified for National Grid, NSTAR, and WMECO separately, resulting in three strata for National Grid and one stratum for both NSTAR and WMECO. Although not specified in the evaluation report, it appears that stratification was based on project size. Figure 24 lists the number of projects and evaluation sample in each stratum by program administrator. Of these projects, five were selected for evaluation to achieve confidence and precision targets of 90% and 10%, respectively, three from National Grid and one each from NSTAR and WMECO.

[100] "Reconciliation Report for DTE Energy's 2010 Energy Optimization Programs." DTE Energy Company. Prepared by Opinion Dynamics Corporation. April 15, 2011.

[101] Examples of ECMs are building envelope upgrades, lighting fixtures and controls, cooling system upgrades, and Energy Management System controls.

[102] Twenty-two custom projects occurred in National Grid service territory, 2 in NSTAR, and 1 in WMECO.

Figure 24. Massachusetts 2008-2010 Custom C&I Sample Summary

Sampling Stratum	Total Number of Projects	Maximum Gross Savings (kWh)	Evaluation Sample
National Grid, 1	12	332,480	1
National Grid, 2	6	608,237	1
National Grid, 3	4	1,108,409	1
NSTAR, 1	2	3,352,840	1
WMECO, 1	1	496,579	1

Source: Navigant Review of Evaluation Report[103]

C.4 Summary from New Mexico (New Mexico Public Service Company and New Mexico Gas Company)

New Mexico Gas Company and the Public Service Company of New Mexico have programs that offer financial incentives to commercial and industrial customers for custom energy efficiency projects.[104] The custom C&I program offered by the New Mexico Gas Company is called "Commercial Solutions" and provides low-flow faucet aerators and pre-rinse spray valves at no cost, as well as a $0.75/therm incentive for custom measures (e.g., water heating, HVAC, building envelope, and industrial process improvements). The custom C&I program offered by the Public Service Company of New Mexico is called the "Commercial Comprehensive Program" and provides rebates for a range of prescriptive and custom measures. Projects are classified as either retrofit, new construction, or QuickSaver direct-install.

The sampling methodology to evaluate C&I programs utilizes stratified random sampling to achieve 90% confidence and 10% precision levels. Projects are stratified by project size. New Mexico Gas Company stratified into three strata. The Public Service Company of New Mexico implemented the sampling strategy for retrofit, new construction, and quick-saver projects separately. Due to the large population of projects for retrofit and QuickSaver, projects were stratified into five strata, while new construction projects were stratified into three strata. Figure 25 and Figure 26 show the number of projects and evaluation sample by stratum.

[103] "Impact Evaluation of 2008 and 2009 Custom CDA Installations." Massachusetts Energy Efficiency Advisory Council. Prepared by KEMA and SBW Consulting Incorporated. June 7, 2011.

[104] El Paso Electric Company also offers a custom C&I program. However, during 2010 and 2011 there were no participants and as a result an evaluation of the program was not conducted.

Figure 25. New Mexico Gas Company 2011 Custom C&I Sample Summary

Sampling Stratum	Total Number of Projects	Evaluation Sample
< 1,000 therms	16	3
1,000 – 5,000 therms	7	3
> 4,000 therms	5	5
Total	**28**	**11**

Source: Navigant Review of Evaluation Report[105]

Figure 26. Public Service Company of New Mexico 2011 Custom C&I Sample Summary

Retrofit		
Sampling Stratum	**Total Number of Projects**	**Evaluation Sample**
< 26.5 MWh	95	5
26.5-50 MWh	38	4
50-150 MWh	48	4
150-500MWh	29	5
>500 MWh	9	9
Total	**224**	**27**

QuickSaver		
Sampling Stratum	**Total Number of Projects**	**Evaluation Sample**
< 10 MWh	192	4
10-20 MWh	150	4
20-40 MWh	88	4
40-95 MWh	44	4
> 95 MWh	10	10
Total	**484**	**26**

New Construction		
Sampling Stratum	**Total Number of Projects**	**Evaluation Sample**
< 70 MWh	12	3
70-250 MWh	9	4
> 250 MWh	2	2
Total	**23**	**9**

Source: Navigant Review of Evaluation Report[106]

C.5 Summary from Pennsylvania (PECO Energy)

The PECO Energy Company Smart Equipment Incentives program offers financial incentives for installing energy-efficient equipment in commercial and industrial facilities and in master-metered multifamily residential buildings. The program offers incentives for both prescriptive and custom measures. Examples of custom projects include energy management systems,

[105]"Evaluation of 2011 DSM Portfolio." New Mexico Gas Company. Prepared by ADM Associates Incorporated. June 2012.

[106]"Evaluation of 2011 DSM & Demand Response Portfolio." Public Service Company of New Mexico. Prepared by ADM Associates Incorporated. March 2012.

compressed air systems, process equipment and chillers, industrial systems, whole building systems, and outdoor lighting. Custom incentive levels are $0.12/kWh for estimated on-peak energy savings and $0.08/kWh for estimated off-peak energy savings, up to 100% of project costs.[107]

In 2010, PECO provided financial incentives to 1,085 non-multi-tenant projects and 490 multi-tenant projects. Of these projects, 39 were selected for evaluation to achieve confidence and precision targets of 85% and 10%, respectively, at the program level.[108] The sample is stratified by project size, based on *ex ante* energy savings, and by project-type (lighting, non-lighting, custom). A three-stage sampling strategy was implemented, with the first stage occurring after the end of Q2, the second stage after Q3, and the third stage after Q4.[109,110] Within the sample, custom projects make up the majority of stratum 1, accounting for 49% of *ex ante* energy savings for the sample population.[111]

C.6 *Summary from Ohio (AEP Ohio)*

AEP Ohio offers commercial and industrial customers energy efficiency incentives through a number of programs. The custom program provides financial incentives for "less common or more complex energy-saving measures" that are installed as part of a qualified retrofit project or equipment replacement project. Examples of custom measures include lighting retrofits, HVAC measures such as VFDs, equipment controls, and process efficiency improvements. Custom incentive levels are based on both energy (kWh) and demand (kW) savings in the first year. Specifically, the incentive levels are $0.08/kWh, $100/kW, up to 50% of the project cost. In 2011, AEP Ohio provided financial incentives to 220 custom projects. Of these, 54 projects were selected for evaluation.

The sampling methodology stratified projects both by geography and by project size. At the time, AEP Ohio had gone through a merger of two regional operating companies so that participants in the custom program were distributed across two rate zone territories. The sample design was conducted separately for each rate zone, targeting confidence and precision levels of 90% and 10%, respectively, for each zone. A two-stage sampling methodology was implemented, with the first wave of projects sampled in November of 2011 and the second wave sampled in February of 2012. Projects were first separated by zone, then stratified based on *ex ante* energy (kWh) savings. Projects were assigned to one of three strata such that there

[107] On-peak hours include 12pm-8pm, June 1 – September 30 (excluding holiday weekdays). Off-peak hours include 8:01pm-11:59am, June 1-September 30, and all hours from October 1-May 31. https://peco.icfi.com/sites/peco/files/2011_PECO_CUSTOM_Incentive_Levels.pdf

[108] The evaluation plan targeted confidence and precision levels of 85% and 15%, respectively. However, the final sample design allowed for 85/10 confidence and precision targets.

[109] The first stage included projects implemented in both Q1 and Q2 due to low levels of participation in the program during Q1.

[110] Note that PECO reports unverified savings quarterly.

[111] Lighting and non-lighting measures account for 19% and 32%, respectively.

was a relatively even distribution of cumulative standard deviation in energy savings between strata. Figure 27 shows the number of total projects and the number of projects in the evaluation sample for each zone and stratum. In total, the evaluation sample represents 62% of *ex ante* gross energy savings for the population.

Figure 27. AEP Ohio 2011 Custom C&I Sample Summary

Sampling Stratum	Total Number of Projects	Evaluation Sample
Zone 1, Stratum 1	5	5
Zone 1, Stratum 2	19	7
Zone 1, Stratum 3	85	12
Zone 2, Stratum 1	8	5
Zone 2, Stratum 2	18	11
Zone 2, Stratum 3	85	14
Total	**220**	**54**

Source: Navigant Review of Evaluation Report[112]

C.7 *Summary from Maryland (covers five Maryland utilities)*

The five EmPOWER Maryland utilities (Baltimore Gas and Electric, Potomac Electric Power Company, Delmarva Power, Southern Maryland Electric Cooperative, and Potomac Edison) offer large commercial and industrial customers financial incentives for the installation of efficiency measures that are complex and/or unique, such as commercial HVAC and industrial process improvements. Baltimore Gas and Electric (BGE) and Southern Maryland Electric Cooperative (SMECO) offer rebates for up to 50% of retrofit projects and up to 75% of the incremental cost of new construction projects. Potomac Electric Power Company (PEPCO) and Delmarva Power (DPL) programs were implemented jointly and offer $0.16/kWh for energy savings in the first year.[113] Potomac Edison (PE) offers $0.05/kWh of *ex ante* energy savings. The target evaluation sample for each utility was 12 projects to achieve confidence and precision levels of 80% and 20%, respectively. At the time the evaluation samples were drawn, only BGE had enough participants to reach the targeted sample of 12. PEPCO/DPL had 10 custom projects completed, SMECO had 7, and PE had 11. For these utilities, the entire population was used as the evaluation sample.[114]

For BGE, the sampling strategy calculated the percentage of population energy (kWh) and demand (kW) savings for each project using equal weights. These percentages were used to sort the population of projects into three strata such that each stratum represented approximately one-third of population savings. Random numbers were then assigned to projects within each

[112] "Program Year 2011 Evaluation Report: Business Custom Program." AEP Ohio. Prepared by Navigant Consulting, Incorporated. May 10, 2012.

[113] As a result, participants in PEPCO and DPL's programs were combined into a single sample.

[114] The final evaluation sample for PEPCO/DPL was reduced to eight due to barriers in doing on-site verification for two custom projects.

stratum. Sample projects from each stratum were selected based on the random number designation. For BGE, the evaluation sample represents 58% of *ex ante* energy savings for the population.

C.8 *Summary from Vermont (Efficiency Vermont)*

Efficiency Vermont offers financial incentives for installing energy-efficient equipment in commercial and industrial facilities as well as multi-family buildings. The evaluation was conducted for two program years, 2007 and 2008. The sample size was chosen to achieve an 80% confidence level and 10% precision level for the entire portfolio of Efficiency Vermont programs.

Sampling occurred in two stages, with the first wave including projects completed by April 30, 2008, and the second wave including projects completed during the remainder of 2008. The sampling methodology categorizes projects by market type (retrofit or new construction/market opportunities) and end use (lighting, HVAC, and other).

The sample of retrofit projects includes projects of all end uses, whereas the evaluation sample of new construction/market opportunities projects only includes lighting projects. Projects were stratified into three strata based on *ex ante* peak demand savings. Because demand reductions are claimed separately for winter and summer, the population of projects/end uses was further stratified by season. In particular, if the estimated peak reduction was higher during winter, projects/end uses were assigned to "winter." If the estimated peak reduction was higher during summer or was roughly equivalent during winter and summer, projects/end uses were assigned to "summer/non-seasonal." Within each stratum, a random number was assigned to each project/end use and ordered. The evaluation sample was then selected from the top of each group. Figure 28 shows the total number of retrofit and NC/MOP projects, as well as the evaluation samples stratified by project size and seasonality.

Figure 28. Efficiency Vermont 2007-2008 Custom C&I Sample Summary

	Total Number of Projects		Evaluation Sample			
Sampling Stratum	**Retrofit**	**NC/MOP**	**Retrofit, Winter**	**Retrofit, Summer**	**NC/MOP, Winter**	**NC/MOP, Summer**
0.8-5 kW	263	652	8	8	15	15
5-35 kW	244	315	16	17	23	26
> 35 kW	64	35	49	49	21	23
Total	**571**	**1,002**	**73**	**74**	**59**	**64**

Source: Navigant Review of Evaluation Report[115]

[115]"Verification of Efficiency Vermont's Energy Efficiency Portfolio for the ISO-NE Forward Capacity Market." Vermont Department of Public Service. Prepared by West Hill Energy and Computing Incorporated. July 29, 2010.