Consistent Methodology for Self-Reported Residential Net-to-Gross Measurement (MA19X03-B-RSRNTG)

Final Report

May 28, 2020

SUBMITTED TO:
Massachusetts Program Administrators and the Energy Efficiency Advisory Council

SUBMITTED BY:
NMR Group, Inc.
Tetra Tech, Inc.

Executive Summary

Background

Study Purpose

The primary purpose of this study was to develop specific questions and algorithms when using the self-report approach for downstream residential net-to-gross (NTG) studies. This study will allow for a consistent measurement of net-to-gross across residential downstream programs and measures.

The NMR team along with an advisory group made up of NTG experts from the residential evaluation team, the PAs, and the EEAC reviewed best practice approaches in self-report net-to-gross studies in the residential sector to develop a recommended approach to future residential net-to-gross studies.

Approach

Methodologies and Guidance Documents Reviewed

- Energy Trust of Oregon Free-Ridership Methodology
- Evaluation Framework for Pennsylvania Act 129 Phase III Energy Efficiency and Conservation Programs (PA Framework)
- Illinois Technical Reference Manual, Version 8.0, Volume 4: Cross-Cutting Measures and Attachments (IL TRM)
- New York Clean Energy Guidance
- California Guidance and more recent impact evaluation conducted for the California Public Utilities Commission
- NEEP Decision-Framework for Determining Net Savings Approach – Supplemental Document #2 to Principles and Guidance, September 2014

Results

Questions and algorithms were developed for each of the following:

- Participant Event Type
- Participant Free-ridership
- Participant Spillover
- Contractor Free-ridership
- Contractor Spillover
- Participant and Contractor Free-ridership Reconciliation

Study Limitations

The main limitation of this study is that these questions and algorithms have not been pretested. They are based on research from other jurisdictions that have been highly vetted. Therefore, the evaluation team, EEAC, and PAs recommend conducting sensitivity analysis after the first few full-scale studies. The study provides recommendations given different scenarios, but the results will dictate when decisions need to be made.

Recommendations

- Utilize the questions and algorithms within this report to allow for consistent reporting of NTG for residential studies.

Table of Contents

ACRONYM GLOSSARY ..... **VII**

EXECUTIVE SUMMARY **1**

BACKGROUND 2

KEY FINDINGS FROM THE LITERATURE REVIEW 2

RECOMMENDED RESIDENTIAL SR NTG QUESTIONS FOR MASSACHUSETTS 4

SECTION 1 INTRODUCTION **6**

1.1 BACKGROUND 6

SECTION 2 METHODOLOGY AND APPROACH **9**

2.1 CONVENE ADVISORY GROUP 9

2.2 CONDUCT LITERATURE REVIEW 9

2.3 DEVELOP SR NTG METHODS AND ALGORITHMS 9

SECTION 3 LITERATURE REVIEW **10**

3.1 CATEGORIZATION OF MASSACHUSETTS RESIDENTIAL PROGRAMS/INITIATIVES 10

3.2 RESIDENTIAL SR NTG APPROACHES 12

3.2.1 Energy Trust Free-Ridership Methodology 16

3.2.2 Evaluation Framework for Pennsylvania Act 129 Phase III Energy Efficiency and Conservation Programs 16

3.2.3 Illinois Statewide Technical Reference Manual for Energy Efficiency, Version 7.0. Volume 4: Cross-Cutting Measures and Attachments (January 1, 2019) . 17

3.2.4 New York Clean Energy Guidance 19

3.2.5 California Guidance Documents 19

3.2.6 NEEP Decision-Framework for Determining Net Savings Approach – Supplemental Document #2 to Principles and Guidance (September 2014) ... 22

3.3 KEY FINDINGS AND CONCLUSIONS FROM THE LITERATURE REVIEW 23

SECTION 4 RECOMMENDED RESIDENTIAL SR NET-TO-GROSS SURVEY QUESTIONS AND ALGORITHMS **24**

4.1 PARTICIPANT MARKET EVENT TYPE 25

4.2 PARTICIPANT FREE-RIDERSHIP 27

4.2.1 Intent Score 27

4.2.2 Influence Score 29

4.2.3 Participant Free-ridership Score 30

4.2.4 Consistency Questions 31

4.3 PARTICIPANT SPILLOVER 32

4.4 CONTRACTOR FREE-RIDERSHIP 34
4.5 PARTICIPATING CONTRACTOR SPILLOVER 36
APPENDIX A ADDENDUM TO LITERATURE REVIEW: SR NTG METHODS AND ALGORITHMS USED IN LEADING STATES 39
A.1 ENERGY TRUST FREE-RIDERSHIP METHODOLOGY 39
A.1.1 Stated Intent in the Absence of the Program / Project Change 40
A.1.2 Program Influence 40
A.1.3 Free Rider Rate 41
A.2 EVALUATION FRAMEWORK FOR PENNSYLVANIA ACT 129 PHASE III ENERGY EFFICIENCY AND CONSERVATION PROGRAMS 41
A.2.1 Free-Ridership for Incentive-Based Programs 42
A.2.2 Free-Ridership for Appliance Recycling Programs 45
A.2.3 Spillover 46
A.2.4 Non-Participant Spillover 48
A.3 ILLINOIS TECHNICAL REFERENCE MANUAL 51
A.3.1 Overview of Residential Survey Based NTG Approaches – IL TRM 51
A.3.2 Participant and Non-Participant Spillover with Customers and Trade Allies 61
A.3.3 Participant and Non-Participant Spillover with Builders (Residential New Construction) 64
A.4 NEW YORK CLEAN ENERGY GUIDANCE 66
A.4.1 Spillover 66
A.4.2 Estimating Net-To-Gross Ratios Using the Self-Report Approach 68
A.4.3 Other NTG approaches 70
A.5 CALIFORNIA GUIDANCE DOCUMENTS 71
A.6 NEEP DECISION-FRAMEWORK FOR DETERMINING NET SAVINGS APPROACH – SUPPLEMENTAL DOCUMENT #2 TO PRINCIPLES AND GUIDANCE, SEPTEMBER 2014 . 77
APPENDIX B ALTERNATIVE APPROACHES FOR ESTIMATING NTG 78
B.1 BILLING DATA ANALYSIS 78
B.2 MARKET SALES DATA ANALYSES 79
B.3 PRICING AND ELASTICITY ANALYSIS 80
B.4 SELF-REPORTED COUNTERFACTUAL 80
B.5 STRUCTURED EXPERT JUDGMENT 81
B.6 HISTORICAL TRACING 81
B.7 TOP-DOWN ECONOMETRIC MODELING 82

APPENDIX C ALGORITHMS FOR THE RECOMMENDED RESIDENTIAL NTG SR SURVEY QUESTIONS 84

C.1 EVENT TYPE ALGORITHM 84

C.2 PARTICIPANT FREE-RIDERSHIP ALGORITHM 85

C.3 PARTICIPANT SPILLOVER ALGORITHM 87

C.4 CONTRACTOR FREE-RIDERSHIP ALGORITHM 88

C.5 PARTICIPANT AND CONTRACTOR FREE-RIDERSHIP RECONCILIATION 89

C.6 PARTICIPATING CONTRACTOR SPILLOVER ALGORITHM 90

Acronym Glossary

Acronym	Full Name
ARRA	American Recovery and Reinvestment Act
BCR	Benefit-cost ratio
CA	California
C&I	Commercial and industrial
CPUC	California Public Utilities Commission
DOE	Department of Energy
DSM	Demand-side management
E	Efficiency
EDC	Electric Distribution Company
EEAC	Energy Efficiency Advisory Council
EM&V	Evaluation, measurement, and verification
EUL	End-of-useful life
ER	Early replacement
ETO	Energy Trust of Oregon
FR	Free-Ridership
HER	Home Energy Report
HES	Home Energy Services
HVAC	Heating, ventilation, and air conditioning
IEPEC	International Energy Program Evaluation Conference
IL	Illinois
ISP	Industry-standard Practice
MA	Massachusetts
ME	Market Effects
MOU	Memorandum of Understanding
NC	New Construction
NEEP	Northeast Energy Efficiency Partnerships
NP	No-program score
NPp	Preliminary no-program score
NPSO	Non-participant spillover
NTG	Net-To-Gross
NTGR	Net-To-Gross Ratio
NY	New York
PA	Pennsylvania
PAs	Program Administrators
PEG	Program Evaluation Group
PI	Program influence score
PSO	Participant spillover
Q	Quantity
ROF	Replace-on-failure
SAG	Stakeholder Advisory Group
SO	Spillover
SR	Self-report

SRA	Self-report approach
SWE	Statewide Evaluator
T	Timing
TRM	Technical Resource Manual
UMP	Uniform Methods Project

ES

Executive Summary

The overarching goal of this study was to develop a consistent methodology for determining self-reported (SR) residential downstream net-to-gross (NTG) for Massachusetts that can be adapted to reflect program offerings where feasible. The NMR evaluation team undertook the following activities to achieve this goal:

- Convene an Advisory Group – The NMR team met periodically with an advisory group to ensure that the team focused on the highest priority issues and obtained feedback and guidance from other NTG experts and users of NTG values. The advisory group comprised NTG experts from the residential evaluation team (NMR and Navigant), the Program Administrators (PAs), and the Energy Efficiency Advisory Council (EEAC).
- Plan and Conduct Literature Review – The team reviewed studies developed as part of the PAs' earlier NTG methodology research and relevant new NTG literature from other jurisdictions that the team and advisory group identified. The team summarized the results of the literature review for the advisory group, identifying the best ideas from these other jurisdictions for consideration and providing a preliminary assessment of the categories or types of programs relevant for consistent SR NTG measurement.
- Identify Residential Programs/Initiatives for which SR NTG is appropriate – To help evaluators determine when to use the SR NTG algorithms presented in this study and to help the PAs to more efficiently prepare their filings, the team created tables aligned with the Benefit-cost Ratio model (BCR) identifying the residential programs/initiatives for which SR NTG is appropriate.
- Develop Methodologies and SR NTG Algorithms for Program Types – Taking into account each of the categories or types of programs and the different groups that would be expected to self-report NTG, the team developed a SR NTG method, including high-level outlines of algorithms. The questions have enough flexibility to address differences in program types.

Section 1 of this report provides background on the study. Section 2 presents the methodology and approach. Section 3 summarizes the literature review conducted in other jurisdictions surrounding self-reports and presents key findings from that review. Section 4 describes the recommended SR NTG questions to be asked of program participants and participating contractors. Appendix A provides more detail on the methods reviewed as part of the literature review. This review was instrumental in developing the recommended SR NTG questions that may serve as useful reference material for future studies. Appendix B summarizes other approaches that could be used to estimate NTG for residential programs, including some of the factors that evaluators should consider regarding each potential NTG approach. Finally, Appendix C presents the algorithms for the recommended SR NTG survey questions.

BACKGROUND

In 2011 for C&I programs[1] and separately for residential programs,[2] the NMR cross-cutting evaluation team conducted net savings methodology studies by reviewing methodologies in use across the nation. These reviews explored the pros and cons of alternative methods for estimating the counterfactual (i.e., what would have happened absent the program) in different contexts.

The C&I downstream NTG studies in Massachusetts have followed a standardized methodology, first established in 2003 and updated in 2011, for use in downstream situations in which end users are able to report on program attribution via self-report methods. The 2011 Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches, or *residential framework*, also recommended the SRA for the Massachusetts residential sector.[3] However, the residential framework did not specify a standard algorithm or set of survey questions. In addition to the SR approach, other net savings methodologies were recommended in the residential framework and have been used since.

The recent Massachusetts report *Net-To-Gross Methodology Research* [4] recommended expanding the residential guidance to include specific questions and algorithms for different program types when using a SR approach. This study fulfills this recommendation.

KEY FINDINGS FROM THE LITERATURE REVIEW

As part of the first meeting with the Advisory Group, the evaluation team, PAs, and EEAC identified additional jurisdictions that warranted research into SR NTG and possible sources, such as specific TRMs and exemplary NTG evaluations in regions that have developed established NTG protocols/methods. As part of the literature review, the evaluation team compiled relevant SR NTG approaches and algorithms from jurisdictions prioritized by the Advisory Group, including leveraging the Massachusetts C&I and Residential NTG frameworks.

The methodologies reviewed include the following:

- Energy Trust of Oregon Free-Ridership Methodology
- Evaluation Framework for Pennsylvania Act 129 Phase III Energy Efficiency and Conservation Programs (PA Framework)
- Illinois Technical Reference Manual, Version 8.0, Volume 4: Cross-Cutting Measures and Attachments (IL TRM)

---

[1] Tetra Tech, KEMA, NMR Group, Inc. (2011). Cross-Cutting (C&I) Free-Ridership and Spillover Methodology Study Final Report. Massachusetts Program Administrators. http://ma-eeac.org/wordpress/wp-content/uploads/Massachusetts-PAs-Cross-Cutting-CI-Free-ridership-and-Spillover-Methodology-Study.pdf

[2] Tetra Tech, NMR Group, Inc., KEMA (2011). Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches. Final report prepared for the Massachusetts Program Administrators. http://ma-eeac.org/wordpress/wp-content/uploads/Cross-Cutting-Net-to-Gross-Methodology-Study-for-Residential-Programs-Suggested-Approaches-Final-Report.pdf

[3] Tetra Tech, NMR Group, Inc., KEMA (2011). Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches. Final report prepared for the Massachusetts Program Administrators.

[4] Tetra Tech (2017). *Net-to-Gross Methodology Research – TXC08.* Prepared for the Program Administrators of Massachusetts.

In addition to these methodologies, we reviewed the following documents, which offered guidance on determining NTG:

- New York Clean Energy Guidance
- California Guidance and more recent impact evaluation conducted for the California Public Utilities Commission
- NEEP Decision-Framework for Determining Net Savings Approach – Supplemental Document #2 to Principles and Guidance, September 2014

The documents reviewed contained three main approaches, as summarized below:

Intent and Influence: The intent element reflects the customer's decision to install the energy-efficient project without program funds. The influence element reflects the influence of the program on the customer's decision to install the energy-efficient project or measure. As noted in the Pennsylvania Framework, "these two components were chosen as a way to balance potentially different and opposing biases."

Timing, Quantity, Efficiency: Variations on this battery ask customers about one or more of the following three elements: timing, quantity, and efficiency. Customers are asked when they would have installed the equipment or measure without the program, how many units they would have installed if the program had not existed, and if the equipment or measure would have been the same level of efficiency or something less efficient.

Appliance Recycling: Most documents detail a separate free-ridership measurement for appliance recycling programs since these programs remove equipment from the home rather than adding or replacing it. As a result, a different battery of free-ridership questions is needed. These questions typically ask customers if, in the absence of the program, the old equipment would have been kept in place, put in a different location in the home, given to another home for use, or discarded and removed from service.

The three primary documents (IL TRM, PA Framework, and the Energy Trust of Oregon) are similar in their approaches for most residential programs, other than appliance recycling, in that they include a consistent set of core questions asking about influence of the program and customer intent had the program not existed. These frameworks have been in place for multiple program cycles, have been well-tested, and are commonly cited approaches. As noted in Appendix A, while these three frameworks vary somewhat in terms of the number of questions included, the questions asked are much shorter than the current MA C&I NTG downstream SR questions. This lower response burden approach recognizes the less-complex decision-making process at the household level.

Some of the more recent SR NTG studies in Massachusetts have incorporated some of the question logic from one or more of these three frameworks. Recent MA residential studies have incorporated influence into NTG algorithms that already have timing, quantity, and efficiency components.

Based on this literature review and discussions with the PAs and EEAC, we recommended building upon the algorithms previously used in Massachusetts and incorporating commonalities from the different frameworks reviewed. Upon approval of this plan by the PAs and EEAC, we did

this, ensuring that questions about influence, timing, efficiency, and quantity, and question(s) designed to ensure consistency, were included in the recommended residential batteries of questions and algorithm(s).

Recommended Residential SR NTG Questions for Massachusetts

The NMR team drafted a streamlined and adaptable approach to the residential NTG batteries of questions for programs where self-reports from participating end-users and contractors are appropriate. These SR batteries apply to downstream programs other than appliance recycling. For appliance recycling, the team responsible for the evaluation should follow the protocols outlined in the Uniform Methods Project.[5]

The NTG batteries consists of the following five groups of questions, which we explain in more detail in Section 4 of this report. Section 4 also contains the detailed questions. The detailed algorithms for these five question batteries can be found in Appendix C of this document.

- **Participant market event type** – used to determine whether the reason for purchasing the equipment was due to new construction, new equipment, replace-on-failure (ROF), or early replacement. This series of questions identifies participants who should be skipped out of the timing question (ROF and new construction).
- **Participant free-ridership** – includes questions on timing, quantity, and efficiency elements; an influence adjustment; and consistency checks. We recommend asking the participant free-ridership questions as soon as possible after receiving the program-eligible equipment to maximize recall, balancing that with the need to use the most recent participants for setting prospective NTG ratios.
- **Participant spillover** – questions ask about whether participants have purchased any additional equipment since participation, the type and efficiency of equipment purchased, the importance of the program on their decision to purchase the additional efficient equipment. A follow-up question asks about the likelihood of purchasing the additional efficient equipment had they not also purchased the rebated equipment. Sufficient time should elapse before asking the spillover questions (e.g., 12 months) to allow the respondent time to have made purchases on their own.
- **Contractor free-ridership** – contains questions to ask contractors/salespersons about FR for measures for which contractors/salespersons play a prominent role in delivering and promoting the equipment. The questions confirm program sales and determine whether the sales would have occurred had the program rebates not been available.
- **Participating Contractor Spillover** – contains questions on the sale of program-eligible equipment that did not receive a rebate through the program, the influence of the program

[5] Keeling, J.; Bruchs, D. (2017). Chapter 7: Refrigerator Recycling Evaluation Protocol. The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68563. http://www.nrel.gov/docs/fy17osti/68563.pdf

on the contractor's recommendations, and the influence of the program and other factors on the number of energy-efficient sales outside the program.

Outlined in its own algorithm, we provide instructions in Appendix C on when participant free-ridership should be used and when contractor free-ridership should replace or be averaged with participant results.

Because these are new questions and new algorithms, the evaluation team, EEAC, and PAs recommend conducting sensitivity analysis after the first few full-scale studies. For example, in the 'Intent' section of the participant FR battery of questions, don't know and refused responses receive a score of 0.5, which is the midpoint of the intent score. The sensitivity analysis would be useful in deciding whether a different score is more appropriate or whether these participants' responses should be removed from the analysis.

Section 1 Introduction

The overarching goal of this study is to develop a consistent methodology for determining self-reported (SR) residential downstream net-to-gross (NTG) for Massachusetts that can be adapted to reflect program offerings where feasible. This final report summarizes the literature review conducted to investigate best practice SR NTG in other jurisdictions and identifies the range of residential downstream program delivery approaches where an SR methodology is and is not recommended. Finally, this report contains recommended survey questions and algorithms to be utilized for program offerings where a SR NTG is feasible.

1.1 Background

Before presenting the key findings of the literature review and the recommended SR NTG batteries, we summarize the history of what has occurred in Massachusetts in the past regarding SR studies.

In 2011 for C&I programs[6] and separately for residential programs,[7] the NMR cross-cutting evaluation team conducted net savings methodology studies by reviewing methodologies in use across the nation. These reviews explored the pros and cons of alternative methods for estimating the counterfactual (i.e., what would have happened absent the program) in different contexts.

The C&I downstream NTG studies in Massachusetts have followed a standardized methodology, first established in 2003 and updated in 2011, for use in downstream situations in which end users are able to report on program attribution via self-report methods. The 2011 methodology study presented survey questions and algorithms for estimating free-ridership (FR) and spillover (SO) using a self-report approach (SRA). This approach consisted of sequential questions used to identify FR and SO. Respondents were asked about their decision-making process and then about the actions they would have taken had the program services not been offered. This approach also assessed the program's impact on project timing, measure quantity, and efficiency levels, while explicitly recognizing that the cost of energy-efficient equipment can be a barrier to installation in the absence of program administrator (PA)-sponsored energy-efficiency programs.

In the most recent Massachusetts Sponsors' Commercial and Industrial Programs Free-ridership and Spillover Study, [8] conducted in 2017, the survey questions and algorithms were modified slightly by the cross-cutting evaluation team to reflect additions and recommendations made in

---

[6] Tetra Tech, KEMA, NMR Group, Inc. (2011). Cross-Cutting (C&I) Free-Ridership and Spillover Methodology Study Final Report. Massachusetts Program Administrators. http://ma-eeac.org/wordpress/wp-content/uploads/Massachusetts-PAs-Cross-Cutting-CI-Free-ridership-and-Spillover-Methodology-Study.pdf

[7] Tetra Tech, NMR Group, Inc., KEMA (2011). Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches. Final report prepared for the Massachusetts Program Administrators. http://ma-eeac.org/wordpress/wp-content/uploads/Cross-Cutting-Net-to-Gross-Methodology-Study-for-Residential-Programs-Suggested-Approaches-Final-Report.pdf

[8] NMR Group, DNV GL, Tetra Tech, Inc (2018). *Massachusetts Sponsors' Commercial and Industrial Programs Free-ridership and Spillover Study*. Final report prepared for the Massachusetts Program Administrators.

the 2012 Residential Heating, Water Heating, and Cooling Equipment evaluation report[9] to distinguish early replacement from replace-on-failure equipment. The evaluation team followed this with a NTG analysis of commercial and industrial customers who had executed a Memorandum of Understanding (MOU) with Eversource.[10] This work explored how MOUs impact program participation, compared MOU NTG survey responses to non-MOU survey responses, and determined whether changes should be considered for the NTG methodology to account for the effect. Interviews with MOU customers showed that they would not have been able to complete the projects they did without the assistance they received from the program through the MOU. Analysis of the NTG survey results of these customers also showed lower free-ridership rates among customers with MOUs. To ensure that the NTG research could account for the impact of an MOU in future research, the study offered three suggestions for modifying future survey wording for MOU customers, as follows:

- Modify the survey questions to explicitly reference the information and assistance they received as part of the MOU when asking about the program assistance they may have received. This would remind customers with MOUs to consider that assistance while responding to the survey questions.
- Add a question that asks the specific respondent about their involvement in the MOU to ensure that the evaluation team is talking with the person directly involved with the MOU.
- The MOU may be a factor in some customers completing energy-efficient projects, but other organizational goals could also have an impact, so add a question regarding the influence of the MOU in the customer's decision to implement the project(s).

The 2017 C&I survey was also revised to assist the C&I Evaluation P73 Global Transition Plan Track C Net-to-Gross Methodology in understanding the impact of an industry-standard practice (ISP) baseline by incorporating questions for three specific technologies[11] that have moved or were moving to an ISP baseline.

The 2011 Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches, or *residential framework*, also recommended the SRA for the Massachusetts residential sector.[12] However, the residential framework did not specify a standard algorithm or set of survey questions. In addition to the SRA, other net savings methodologies were recommended in the residential framework and have been used since. For example, the net impact evaluation of the Residential Behavior Program's Home Energy Reports used self-reports and a billing data analysis to arrive at net savings.[13] The Home Energy Services (HES) NTG evaluation used participant and non-participant self-reports, as well as discrete choice modeling

---

[9] Navigant, Opinion Dynamics, Cadmus (2013). *2012 Residential Heating, Water Heating, and Cooling Equipment Evaluation: Net-to-Gross, Market Effects, and Equipment Replacement Timing*. Final report prepared for the Massachusetts Program Administrators.

[10] Tetra Tech & NMR Group, Inc. (2019). "TXC49 (C&I NTG) MOU Research Results." Memorandum prepared for the Massachusetts Program Administrators.

[11] Air compressors, condensing boilers, and LED lighting.

[12] Tetra Tech, NMR Group, Inc., KEMA (2011). Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches. Final report prepared for the Massachusetts Program Administrators.

[13] Navigant and Illume (2015). *Massachusetts Cross-Cutting Behavioral Program Evaluation OPower Results.* Final report prepared for the Massachusetts Program Administrators.

to estimate the impacts of the HES program's marketing and incentives on consumer decisions.[14] A net savings study of the Residential New Construction program relied on structured expert judgment.[15]

[14] Cadmus and Navigant (2012). *Home Energy Services Net-to-Gross Evaluation.* Final report prepared for the Massachusetts Program Administrators.

[15] NMR Group, Inc. (2014). *Residential New Construction Net Impacts Report.* Final report prepared for the Massachusetts Program Administrators.

Section 2 Methodology and Approach

This study is based on a literature review and a review of the gas and electric BCR models, described below.

2.1 Convene Advisory Group

The NMR team met periodically with an advisory group to ensure that the team focused on the highest priority issues and obtained feedback and guidance from other NTG experts and users of NTG values. The advisory group comprised NTG experts from the residential evaluation team (NMR and Navigant), the PAs, and the EEAC.

2.2 Conduct Literature Review

For the literature review, the team reviewed studies developed as part of the PAs' earlier NTG methodology research and relevant new NTG literature that the team and advisory group identified. The team summarized the results of the literature review for the advisory group, identifying the best ideas from other jurisdictions for consideration and providing a preliminary assessment of the categories or types of programs relevant for consistent SR NTG measurement. Section 3 presents more detail about the literature included in the review.

The team also reviewed the PAs' program plans to identify the residential programs/initiatives to address in the SR NTG research and determine if different algorithms or questions would be needed for certain programs. To guide evaluators and help PAs prepare their filings more efficiently, the team created tables showing the residential program delivery mechanisms by measure for each of the combinations of residential programs, initiatives, and sub-offerings in the benefit-cost ratio model (BCR).

2.3 Develop SR NTG Methods and Algorithms

Based on the literature review and input from the advisory group, the team identified methods and algorithms for gathering SR NTG data and calculating NTG for selected types of programs and situations. The team presented the draft methods and algorithms to the advisory group and refined these based on the advisory group's feedback.

More detail on the approach used for the literature review and key findings from the literature review are presented in Section 3. Recommendations for the survey questions to include in the methodology for consistent NTG measurement in Massachusetts for residential programs using a SR approach can be found in Section 4. More details on the approaches examined as part of the literature review, along with survey questions and algorithms, can be found in Appendix A. Appendix B summarizes other approaches that could be used to estimate NTG for residential programs, including some of the factors that evaluators should consider regarding each potential NTG approach. Appendix C presents the algorithms for the recommended survey questions.

Section 3 Literature Review

3.1 Categorization of Massachusetts Residential Programs/Initiatives

In conjunction with the literature review, the team created and provided PAs with tables aligned with the BCR identifying the residential programs/initiatives for which SR NTG is appropriate. The tables present this information by program, initiative, sub-offering, end-use, measure/product, delivery mechanism, and measure category (an aggregation of measures). The structure of the tables is meant to help evaluators determine when to use the SR NTG algorithms presented in this study and to help the PAs to prepare their filings more efficiently. As part of the first meeting with the Advisory Group, the evaluation team, PAs, and EEAC identified additional jurisdictions that warranted research into SR NTG and possible sources, such as specific TRMs and exemplary NTG evaluations in regions with established NTG protocols/methods. The Advisory Group decided that, at a minimum, these regions would include Illinois, Energy Trust of Oregon (ETO), and Pennsylvania due to all three's efforts to streamline approaches and standardize methodologies utilizing the SRA approach.

This study builds on the recent Massachusetts SRA approaches methods review study conducted as part of the study Net-to-Gross Methodology Research (TXC08).[16] The TXC08 study reviewed the general methods used in Massachusetts for estimating net savings and compared them to methods used in other regions across the country. The current literature review builds on this research where the evaluation team compiled relevant SR NTG approaches and algorithms from jurisdictions prioritized by the Advisory Group, including leveraging the Massachusetts C&I and Residential NTG frameworks.

The methodologies reviewed include the following:

- Energy Trust of Oregon Free-Ridership Methodology[17]
- Evaluation Framework for Pennsylvania Act 129 Phase III Energy Efficiency and Conservation Programs[18]
- Illinois Technical Reference Manual, Version 8.0, Volume 4: Cross-Cutting Measures and Attachments[19]

---

[16] Net-to-Gross Methodology Research – TXC08, prepared for the Program Administrators of Massachusetts, March 24, 2017. http://ma-eeac.org/wordpress/wp-content/uploads/Net-to-Gross-Methodology-Research.pdf

[17] Energy Trust Free-Ridership Methodology, prepared by Phil Degens, Sarah Castor, and Erika Kociolek, Updated August 7, 2013. https://www.energytrust.org/wp-content/uploads/2016/12/Energy_Trust_Free_Ridership_Methods.pdf

[18] Evaluation Framework for Pennsylvania Act 129 Phase III Energy Efficiency and Conservation Programs, Prepared by NMR Group, Inc., Ecometric Consulting, LLC, Demand Side Analytics, LLC, August 25, 2016. http://www.puc.pa.gov/Electric/pdf/Act129/SWE_PhaseIII-Evaluation_Framework082516.pdf f

[19] Illinois Statewide Technical Reference Manual for Energy Efficiency, Version 8.0. Volume 4: Cross-Cutting Measures and Attachments. October 17, 2019. https://s3.amazonaws.com/ilsag/IL-TRM_Effective_01-01-20_v8.0_Vol_4_X-Cutting_Measures_and_Attach_10-17-19_Final.pdf

In addition to these methodologies, we reviewed the following documents which offer guidance on determining NTG:

- New York Clean Energy Guidance[20]
- California Guidance[21] and more recent impact evaluation conducted for the California Public Utilities Commission[22]
- Northeast Energy Efficiency Partnerships (NEEP) Decision-Framework for Determining Net Savings Approach – Supplemental Document #2 to Principles and Guidance, September 2014[23]

The NEEP Decision Framework document summarizes three literature reviews and white papers on measuring net savings.[24] This white paper briefly describes each of the net savings approaches summarized in the Uniform Methods Project (UMP),[25] highlights the ways these approaches could be used to estimate key components of net savings and discusses the strengths and weaknesses of each method. The NEEP document does not contain any recommended survey questions or algorithms for estimating net savings.

The evaluation team also reviewed several proceedings from the International Energy Program Evaluation (IEPEC) conferences, dating back to 2013, that included topics such as NTG savings attributions, spillover, and program influence. Our initial pull of documents was from both the C&I and residential sectors, as we felt some findings might apply to both sectors. The proceedings we found focused on specific program types (i.e., finance or custom programs) rather than NTG survey questions or algorithms. These proceedings discussed alternative approaches (saturation studies or site visits) and identified steps program implementers could take to increase program influence but did not address the topic of a standardized batteries of questions and algorithm(s) to measure free-ridership in the residential sector.

---

[20] Evaluation, Measurement & Verification Guidance, 2016-11-01. http://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/255ea3546df802b585257e38005460f9/$FILE/CE-05-EMV%20Guidance%20Final%20%2011-1-2016.pdf

[21] The California Evaluation Framework, (2004), Prepared for the California Public Utilities Commission and the Project Advisory Group. https://www.cpuc.ca.gov/uploadedFiles/CPUC_Public_Website/Content/Utilities_and_Industries/Energy/Energy_Programs/Demand_Side_Management/EE_and_Energy_Savings_Assist/CAEvaluationFramework.pdf

[22] Impact Evaluation Report, Home Upgrade Program – Residential Program Year 2017, prepared for the California Public Utilities Commission, April 29, 2019. http://www.calmac.org/publications/CPUC_GroupA_Res_PY2017_HUP_toCALMAC.pdf

[23] Decision-Framework for Determining Net Savings Approach – Supplemental Document #2 to Principles and Guidance "Gross and Net Savings: Principles and Guidance", prepared for the Northeast Energy Efficiency Partnerships, September 2014. https://neep.org/gross-and-net-savings-principles-and-guidance

[24] The three studies summarized in the principles and guidance are as follows: NMR Group, Tetra Tech & KEMA 2011, "Massachusetts Program Administrators Cross-Cutting Net to Gross Methodology Study for Residential Programs –Suggested Approaches (Final), July; Violette D. & Rathbun, P. 2014, "Chapter 17: Estimating Net Energy Savings: Common Practices," Uniform Methods Project, September; and Tetra Tech, KEMA & NMR 2011, "Massachusetts Program Administrators Cross-Cutting C&I Free-Ridership and Spillover Methodology Study Final Report".

[25] Chapter 21: Estimating Net Savings – Common Practices: Methods for Determining Energy-Efficiency Savings for Specific Measures", 2017, prepared for the National Renewable Energy Laboratory. , https://www.nrel.gov/docs/fy17osti/68578.pdf

The list of IEPEC papers the evaluation team reviewed includes the following:

- "The Estimation of Spillover: EM&V's Orphan Gets a Home," 2013 International Energy Program Evaluation Conference, Chicago, Prahl & Associates, Ridge & Associates, TecMarket Works, New York Department of Public Service[26]
- "Not all Spillover is the Same – So Don't Treat it That Way!," 2017 International Energy Program Evaluation Conference, Baltimore, MD, Research Into Action, Inc., ADM Associates Inc.[27]
- "Spillover – Worth Crying Over: An Innovative Methodology for Quantifying Non-Participant Spillover in California," 2015 International Energy Program Evaluation Conference, Long Beach, Itron, CPUC, Energy Division, Opinion Dynamics[28]
- "Telling the "Story" of Program Influence for Custom Programs in CA," 2015 International Energy Program Evaluation Conference, Long Beach, Itron, Inc.[29]
- "Financing or Incentives: Disentangling Attribution," 2017 International Energy Program Evaluation Conference, Baltimore, MD, Opinion Dynamics Corporation[30]
- "Show Me the Attribution: Recommended Methods for Estimating Impacts for Finance Programs," 2015 International Energy Program Evaluation Conference, Long Beach, Opinion Dynamics Inc.[31]
- "How Much Do We Know About Savings Attribution to a Program?," 2017 International Energy Program Evaluation Conference, Baltimore, MD, Opinion Dynamics[32]
- "Gross Is Gross and Net is Net: Simple, Right?," 2013 International Energy Program Evaluation Conference, Chicago, Ridge & Associates, SBW Consulting, TecMarket Works, Prahl & Associates, New York State Department of Public Service[33]

3.2 Residential SR NTG Approaches

In Table 1, we present a high-level summary of the documents the evaluation team reviewed and how NTG elements are measured within each jurisdiction. The table is followed by a discussion of the NTG measurement approaches in each jurisdiction.

The documents reviewed contained three main approaches.

Intent and Influence: The intent element reflects the customer's decision to install the energy-efficient project without program funds. The influence element reflects the influence of the program on the customer's decision to install the energy-efficient project or measure. As noted in

[26] https://www.iepec.org/conf-docs/conf-by-year/2013-Chicago/095.pdf#page=1
[27] http://www.iepec.org/2017-proceedings/65243-iepec-1.3717521/t001-1.3718144/f001-1.3718145/a001-1.3718258/an003-1.3718263.html
[28] https://www.iepec.org/wp-content/uploads/2015/papers/131.pdf
[29] https://www.iepec.org/wp-content/uploads/2015/papers/056.pdf
[30] http://www.iepec.org/2017-proceedings/65243-iepec-1.3717521/t001-1.3718144/f001-1.3718145/a001-1.3718258/an004-1.3718261.html
[31] https://www.iepec.org/wp-content/uploads/2015/papers/111.pdf
[32] https://www.iepec.org/wp-content/uploads/2018/04/2017paper_wayland_patterson_randazzo.pdf
[33] https://www.iepec.org/conf-docs/conf-by-year/2013-Chicago/094.pdf#page=1

the Pennsylvania Framework, “these two components were chosen as a way to balance potentially different and opposing biases.”

Timing, Quantity, Efficiency: Variations on this battery ask customers about one or more of the following three elements: timing, quantity, and efficiency. Customers are asked when they would have installed the equipment or measure without the program, how many units they would have installed if the program had not existed, and if the equipment or measure would have been the same level of efficiency or something less efficient.

Appliance Recycling: Most documents detail a separate free-ridership measurement for appliance recycling programs since these programs remove equipment from the home rather than adding or replacing it. As a result, a different battery of free-ridership questions is needed. These questions typically ask customers if, in the absence of the program, the old equipment would have been kept in place, put in a different location in the home, given to another home for use, or discarded and removed from service.

Below, we summarize the SR approaches presented in the literature.

Table 1: Summary of SR Approaches in Documents Reviewed

Element	Energy Trust of Oregon	Illinois TRM	PA Framework	New York Clean Energy Guidance	CA Guidance	NEEP Framework
Free-ridership measurement	Influence and intent	Two primary approaches: timing, efficiency, and quantity with a program influence adjustment, and a separate battery for appliance recycling	Two approaches: influence and intent, and a separate battery for appliance recycling	Offers guidance on best practices	Timing, Efficiency, and Likelihood	Discusses different options to measure FR
Spillover measurement	No	Participant and non-participant spillover; non-participant spillover can be conducted through non-participant research or trade ally research	Participant, non-participant is optional; non-participant spillover captured through a general population (non-participant) survey or a survey of trade allies	Participant and non-participant spillover	No	Participant and non-participant spillover
Questions and algorithms	Yes	Yes	Yes	No	Yes	No
Timing FR/SO studies	Surveys occur approximately one month after receipt of incentive check	NA	NA	Recommends conducting FR as soon as possible after participation; spillover conducted later	As soon after installation as possible	Citing the UMP Net Savings document; FR data collection should occur soon after participation and spillover should occur later

Element	Energy Trust of Oregon	Illinois TRM	PA Framework	New York Clean Energy Guidance	CA Guidance	NEEP Framework
Frequency of NTG studies	NA	Typically, once per four-year cycle	Conducted at least once per program during every program cycle or phase	NA	NA	NA
Early Replacement (ER) vs. ROF	Not discussed	Uses a single algorithm in the NTG battery; timing is asked about in each algorithm, but the baseline adjustments due to ROF and ER are handled under the gross savings determination	Is addressed in verified savings calculations	Provides guidance on the types of questions that should be asked for estimating gross savings due to ROF and ER	NA	Not discussed
Consistency questions	Discusses how inconsistencies can occur but does not adjust	Recommends asking consistency questions	Not recommended	Recommends asking consistency questions	Recommends asking consistency questions	Not mentioned

3.2.1 Energy Trust Free-Ridership Methodology

The bullets below summarize the historic NTG measurement for Energy Trust of Oregon. Per Phil Degens from Energy Trust of Oregon, NTG does not apply after December 31, 2019. The Energy Trust of Oregon still plans to look at program influence after 2019, but the focus will be on program redesign or planning purposes.

- One goal of this methodology was the ability to apply it to all programs and markets.
- An additional goal was to obtain SR results through a reduced set of survey questions. The thought was that having a short battery would increase response rates.
- The NTG is calculated as 1- free-ridership.
- With this approach, a project's free-ridership score is composed of two elements: a stated intent score and an influence score.
- Spillover is not included in the Energy Trust of Oregon's methodology. Given that the survey is administered close to program participation, it allows little time for spillover to occur.
- The methodology discusses how inconsistent responses can happen and examples of such responses, but the question algorithm does not adjust for consistency checks.
- The surveys are administered approximately one month after receipt of the program incentive on a rolling basis as part of the Energy Trust of Oregon's Fast Feedback survey. The methodology noted that talking with customers soon after participation would result in more accurate data due to reduced recall bias.
- There is no discussion on methodology adjustments based on early replacement or replacing on failure.

3.2.2 Evaluation Framework for Pennsylvania Act 129 Phase III Energy Efficiency and Conservation Programs

The Statewide Evaluator (SWE) determined that the SRA should be used to estimate free-ridership and spillover for downstream programs.

- NTG is captured as NTG = 1 – FR + SO + market effects (ME). Care needs to be taken to ensure there is no double-counting of savings between SO and ME.
- The battery measures two components of free-ridership: (1) *intention* to carry out the energy-efficient project without program funds and (2) *influence* of the program in the decision to carry out the energy-efficient project. The Framework notes that these two components were chosen as a way to balance potentially different and opposing biases – the intention component typically indicates higher free-ridership than the *influence* component. The logic is that combining the two components decreases the potential biases.
- There are two protocols: one for incentive-based programs and one for appliance recycling.

  - For incentive-based programs, the survey questions assess participant intention and program influence. The PA Framework discusses alternative wording for programs that do not fit the standard retrofit incentive model (e.g., direct install, financing).
  - The Appliance Recycling approach follows the approach outlined for refrigerator recycling in the UMP chapter.[34]
- The PA Framework notes that the estimation of participant spillover should assess the number and description of non-incented energy-efficiency measures implemented since program participation, the energy savings associated with those energy-efficiency measures, and the program's influence on the participant's decision to implement the identified measures. Estimation of non-participant spillover is optional and can be done through a general population survey or a survey of trade allies.
- The PA Framework recommends not using consistency checks because it may not result in more accurate data, it may impact later responses to the survey, and it lengthens the survey. Further, the need to determine which response is correct brings more evaluator subjectivity into the assessment.
- Conducted at least once per program during every program cycle or phase.
- Event type is addressed in verified savings calculations, leveraging TRM assumptions for baseline equipment.

3.2.3 Illinois Statewide Technical Reference Manual for Energy Efficiency, Version 7.0. Volume 4: Cross-Cutting Measures and Attachments (January 1, 2019)

The Illinois Statewide Technical Reference Manual for Energy Efficiency (IL TRM) specifies that most programs, other than behavioral programs, use an SRA with customers and trade allies.

- The residential protocol formulates the core NTG as 1- free-ridership + participant spillover + non-participant spillover and provides specific questions and scoring algorithms for measuring each. The free-ridership questions vary depending upon the specific program protocols.
  - The Prescriptive Rebate (with no audit) protocol includes questions about program influence and intent (referred to as the no-program score), as well as consistency checks. There are two different approaches: a basic and enhanced approach. The basic approach only includes questions for customers. An enhanced method may utilize trade ally surveys to provide another quantitative assessment, which may be triangulated with the basic method approach. While the TRM does not specify when an enhanced approach should be used, it does say that for programs where trade allies play a prominent role in delivering the energy-efficiency measure and promoting the program, an estimate of free-ridership from trade allies can be combined with one from participants to form a combined free-ridership value.

[34] "Chapter 7: Refrigerator Recycling Evaluation Protocol" (2017). Developed for the National Renewable Energy Laboratory. https://www.nrel.gov/docs/fy17osti/68563.pdf

- The Single-Family Home Energy Audit protocol uses two different approaches depending upon whether no-cost direct install measures or rebated/discounted measures are included. The no-cost approach asks a simplified set of questions on timing, efficiency, and quantity. The rebated/discounted approach is identical to the Prescriptive Rebate (with no audit) protocol.
- The Multifamily protocol consists of surveys with tenants or with property managers or owners (for common area and building shell improvements). The approach for discounted measures is identical to that used in the Prescriptive Rebates (with no audit) protocol.
- The Energy Saving Kits and Elementary Education protocol is similar to the no-cost direct install approach with questions on timing, efficiency, and quantity.
- The Residential New Construction protocol is based on surveys with builders. The questions ask about program influence, and whether builders planned to build homes to the same standard before learning of the program.
- The Appliance Recycling protocol follows the UMP protocol for residential refrigerator recycling.[35] This approach uses a multi-step process to segment participants into different groups (appliance would have been kept, appliance would have been transferred to another customer for continued use, or appliance would have been discarded and permanently removed from service). This can be enhanced by using additional research efforts such as a retailer survey, an appliance market assessment study, and/or a non-participant survey.
- The IL TRM has a general approach to measuring participant and non-participant spillover from customers. Some individual protocols have alternative questions for measuring spillover from customers, trade allies, and builders.
- Open-ended questions for consistency checks are to be included to check the consistency of responses associated with the core FR methodology and seek clarification on how the program influenced decision-making. Evaluators will review responses and use expert judgment to adjust scores.
- The IL TRM specifies surveying customers from a random sample of current participants, or up to one year of previous program participants, for determining NTG at least once per program plan cycle. The TRM further specifies that the NTG calculation should consider spillover where feasible. Spillover should be measured within the 12 months before the survey date. At a minimum, a sector- or portfolio-level spillover analysis should be considered at least once every Plan period.
- Early replacement determination is based on whether the existing unit was operating when replaced or if the existing unit required repairs. This adjustment is made as part of gross savings baseline determination.

[35] "Chapter 7: Refrigerator Recycling Evaluation Protocol" (2017). Developed for the National Renewable Energy Laboratory. https://www.nrel.gov/docs/fy17osti/68563.pdf

3.2.4 New York Clean Energy Guidance

The NY Clean Energy Guidance (the NY Guidance) does not specify questions or algorithms but does provide guidance on best practices for evaluators to use in an SRA. The NY Guidance also discusses approaches that evaluators can use other than the SR approach.

- NTG is calculated as 1 – FR + SO.
- Spillover can be collected from both participants and non-participants.
- The NY Guidance states that an evaluation using SR methods should include checks for consistency. The evaluator should describe the methods chosen, as well as the rationales for using or not using the techniques for mitigating inconsistencies.
- In order to minimize the problem of recall, SRA interviews addressing FR should be conducted with the decision maker(s) as soon after the installation of equipment as possible. Interviews or other data collection to assess spillover need to be conducted later to allow enough time for the occurrence of spillover.
- The NY Guidance provides a series of questions to identify early replacement versus replace on failure, which are used in estimating gross savings.

3.2.5 California Guidance Documents

3.2.5.1 California Evaluation Framework

The California Evaluation Framework (CA Framework), developed in 2004 for the California Public Utilities Commission (CPUC) Energy Division and the Project Advisory Group[36], provides guidance on different methods that can be used to assess free-ridership, including the SRA.

- The CA Framework notes that the SRA involves asking one or more key participant decision-makers a series of closed and open-ended questions about their motivations for installing the efficient equipment and the efficiency of the equipment that they would have installed in the absence of the program. Additional questions are typically asked to establish the temporal precedence of the program and to rule out rival explanations for the installation.
- The CA Framework states that adding a consistency check question and adjusting an individual's estimate accordingly also improves the estimate. The CA Framework recommends dropping respondents with inconsistent responses rather than assuming central estimates for them.
- The CA Framework recommends developing a standardized survey method for use in California, noting that a carefully constructed survey with checks and triangulation methods might offer a more *accurate* measurement of a program's free-ridership.

[36] The California Evaluation Framework, (2004), Prepared for the California Public Utilities Commission and the Project Advisory Group.[36]
https://www.cpuc.ca.gov/uploadedFiles/CPUC_Public_Website/Content/Utilities_and_Industries/Energy/Energy_Programs/Demand_Side_Management/EE_and_Energy_Savings_Assist/CAEvaluationFramework.pdf

In terms of timing, the CA Framework notes that the best available estimate of the NTGR is the latest estimate for a program. The Framework also recommends that deemed NTGR should never be used to report net savings within an evaluation.

In addition to the CA Framework, guidelines have also been established for estimating NTG ratios using the SRAs.[37] The SRAs deviate from the standard approach to assessing causality (experiment or quasi-experiment), which is not always desirable or possible. The SRA guidelines are as follows:

1. Interview customers as soon after the installation as possible.
2. Identify the correct respondent(s).
3. Use set-up questions to establish the context and sequence of events that led to decisions.
4. Use multiple questions, both quantitative and qualitative, to increase reliability.
5. Assess validity and reliability for each question used.
6. Use consistency checks for handling inconsistent responses while the respondent is on the phone.
7. Ask measure-specific questions.
8. Explore cases where participants would have installed something more efficient than the program assumed baseline but not as efficient as program equipment (partial FRs).
9. Measure the program's impact on the acceleration of installation (deferred FR).
10. Document the scoring algorithms and test using sensitivity analyses. A preponderance of evidence approach is better than relying solely on an algorithm.
11. Determine how to handle "don't know" and non-response responses.
12. Weight the data to take account for the size of the savings impacts at the customer or project level.
13. Ask open-ended questions regarding other possible reasons for installing the efficient equipment.
14. Report precision levels.
15. Pretest the instrument.
16. Incorporate additional quantitative and qualitative data for large projects or complex decision making:
17. Use multiple respondents.
    a. Use other site- and market-level data.
    b. Establish rules for data integration.
    c. Analysis method (e.g., case studies are one method for assessing both quantitative

[37] Ridge, Richard, Ken Keating, Lori Megdal, and Nick Hall. *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches.* October 15, 2007.

and qualitative data). Use multiple evaluators to independently review the data.

18. Train engineers to collect the data for complex technologies.

3.2.5.2 Residential Retrofit High Impact Measure Evaluation Report

The CPUC Energy Division convened a committee of experienced evaluators to develop a standard framework for the measurement and calculation of NTG ratios for residential and small commercial programs in a systematic and consistent manner using the SRA approach.[38] The committee designed the approach to fully comply with the *Evaluator Protocols* and the G*uidelines for Estimating Net-To-Gross Ratios (NTGR) Using the Self-Report Approaches.*

The committee's goal was to develop a consistent set of survey questions with specific fill-in areas where the questions could be customized for the program being evaluated while maintaining consistency across the portfolio. The questions proposed by the committee for the free-ridership battery are contained in Appendix A of the Residential Retrofit Hight Impact Measure Evaluation Report. Customized surveys by program are also contained in the appendix.

The residential/small commercial NTGR algorithm derived four separate measurements of free-ridership:

1. The first consisted of responses to a series of yes/no questions that measured the impact of the program on the quantity, efficiency, and timing of the purchase.
2. The second consisted of a 0 to 10 scale that asked the likelihood that the respondent would have purchased the same exact high-efficiency measure in the absence of the program.
3. The third measurement combined responses to the quantity and timing questions with responses on a 0-to 10-scale to a question that asked the respondent's agreement with the statement that, in the absence of the program, they would have paid the additional rebate amount to buy the high-efficiency equipment on their own.
4. The final measurement combined responses to the quantity and timing questions with responses on a 0 to 10 scale to a question that asked the respondent's agreement with the statement that the program was a critical factor in their decision to purchase the high-efficiency equipment.

These questions covered all the requirements provided in the *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches*, such as multiple questions, efficiency level, likelihood of adoption, timing and quantity, and consistency checks.

When responses were inconsistent among the four measurements contained in the residential/small commercial NTGR algorithm, the guidelines prescribe that an analyst review the

[38] The Residential Retrofit High Impact Measure Evaluation Report and Appendices, (February 8, 2010, Prepared for the California Public Utilities Commission Energy Division) contains the questions used in the residential battery. http://www.calmac.org/publications/FinalResidentialRetroEvaluationReport_11.pdf http://www.calmac.org/publications/FinalResidentialRetroEvaluationAppendices.pdf

responses to open-ended questions that asked for clarification of the inconsistency and recode the four measurements as needed. The four separate measurements are then averaged to develop a final free-ridership estimate at the measure level. Before finalizing the NTGR algorithm, the committee conducted iterative testing with a partial data set. This testing contributed to the reliability of the algorithm and its computer coding.

The battery developed by the Residential NTG Committee is long and asks several very detailed questions. A more recent impact evaluation conducted for the CPUC for the Home Upgrade Program[39] used a more streamlined battery of questions to assess the extent of free-ridership in the program. The participant survey followed CPUC guidelines to assess free-ridership based on SR responses, considering respondent fatigue, complexity, timing, and budget constraints. The survey was finalized based on input from the PAs.

The Home Upgrade Program study defined full free-riders as those who would have installed exactly the same measure at the same quantity (Q), efficiency (E), and time (T) in the absence of the program. The survey captured both full and partial free riders who would have undertaken or installed the measure(s), but of lesser quantity, and/or at lesser efficiency, or a different time. Respondents selected one of two options when they began the survey: whether they (1) considered the project as one buying decision or (2) considered each measure installed as a separate buying decision. If they considered the project as one buying decision, they received a short-form battery that applied to the entire project. If they considered each measure installed as a separate buying decision, they were given a long-form battery for each measure they reported installing. Those who received the short form battery were not asked the quantity question since they considered the whole project as one single decision.

3.2.6 NEEP Decision-Framework for Determining Net Savings Approach – Supplemental Document #2 to Principles and Guidance (September 2014)

This document summarizes three literature reviews and white papers on measuring net savings. It highlights the different methods that can be used to estimate net savings under different conditions, such as custom measures, measures with few and diverse participants, measures with large numbers of similar participants, and measures with upstream influence. It rates the typical cost or complexity of each method, as well as any special requirements. This document also groups the various approaches on a continuum, identifying those techniques that are driven heavily by data compared to those techniques that are driven heavily by judgement. This document does not recommend specific survey questions or algorithms.

[39] Impact Evaluation Report, Home Upgrade Program – Residential Program Year 2017, prepared for the California Public Utilities Commission, April 29, 2019.
http://www.calmac.org/publications/CPUC_GroupA_Res_PY2017_HUP_toCALMAC.pdf

3.3 Key Findings and Conclusions from the Literature Review

The three primary documents (IL TRM, PA Framework, and the Energy Trust of Oregon) are similar in their approach for most residential programs, other than appliance recycling, in that they include a consistent set of core questions asking about influence of the program and customer intent if the program did not exist. (See Appendix A for the survey questions.) These frameworks have been in place for multiple program cycles, have been well-tested, and are commonly cited approaches. As noted in Appendix A, while these three frameworks vary somewhat in terms of the number of questions included, the number of questions is much shorter than the current MA C&I NTG downstream battery. This lower response burden recognizes the less-complex decision-making process at the household-level.

The Energy Trust of Oregon recommends conducting the survey about one month after receipt of the program incentive. The documents from other jurisdictions that specify survey timing recommended conducting FR surveys as soon as possible after program participation.

Some of the more recent SR NTG studies in Massachusetts have incorporated some of the question logic from one or more of these three frameworks, such as adding a question to capture program influence and limiting the number of questions in the NTG battery. Recent MA residential studies have incorporated influence into NTG algorithms that already have timing, quantity, and efficiency components.

Based on this literature review and discussions with the PAs and EEAC, the team recommended building upon the algorithms previously used in Massachusetts and incorporating commonalities from the different frameworks reviewed. We ensured that questions about influence, timing, efficiency, and quantity, and question(s) designed to ensure consistency, were included in our recommended residential batteries of questions and algorithm(s). These batteries are presented in Section 4 and Appendix C of this report.

Section 4 Recommended Residential SR Net-to-Gross Survey Questions and Algorithms

The evaluation team drafted a streamlined and adaptable approach to develop batteries of questions for residential NTG studies of programs for which self-reports from participating end-users or contractors are appropriate. The batteries described in this section apply to downstream programs other than appliance recycling. For appliance recycling, the team responsible for the evaluation should follow the protocols outlined in the Uniform Methods Project.[40]

The NTG batteries consist of the following five different groups of questions, which are explained in more detail in the remainder of this chapter. The more detailed algorithms for these five question batteries can be found in Appendix C of this document.

- **Participant market event type** – used to determine whether the reason for purchasing the equipment was due to new construction, the addition of new equipment (to an existing home), replace on failure (ROF), or early replacement. This series of questions identifies participants who should be skipped out of the timing question (ROF and new construction).
- **Participant free-ridership** – includes questions on timing, quantity, and efficiency elements; an influence adjustment; and consistency checks. We recommend asking the participant free-ridership questions as soon as possible after receiving the program-eligible equipment to maximize recall, balancing that with the need to use the most recent participants for setting prospective NTG ratios.
- **Participant spillover** – questions ask about whether participants have purchased similar equipment since participation and the type and efficiency of equipment purchased. Respondents are then asked about the importance of the program on their decision to purchase the additional efficient equipment and the likelihood of purchasing the additional efficient equipment had they not also purchased the rebated equipment. A sufficient period should elapse before asking the spillover questions (e.g., 12 months) to allow the respondent time to have made purchases on their own.
- **Contractor free-ridership** – contains questions to ask participating contractors/ salespersons about FR for measures for which contractors/salespersons play a prominent role in delivering and promoting the equipment. The questions confirm program sales and determine whether the sales would have occurred had the program rebates not been available.
- **Participating Contractor Spillover** – contains questions on the sale of program-eligible equipment that did not receive a rebate through the program, the influence of the

[40] Keeling, J.; Bruchs, D. (2017). Chapter 7: Refrigerator Recycling Evaluation Protocol. The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures. Golden, CO; National Renewable Energy Laboratory. NREL/SR-7A40-68563. http://www.nrel.gov/docs/fy17osti/68563.pdf

program on the contractor's recommendations, and the influence of the program and other factors on the number of energy-efficient sales outside the program.

Together, these algorithms are used to calculate an NTG ratio, as follows:

NTG = (1 – FR) + participant spillover + contractor spillover.

The free-ridership component is taken from the participant, the contractor, or a combination of the two. We provide instructions in Appendix C on when participant free-ridership should be used and when contractor free-ridership should replace or be averaged with participant results.

CAVEAT

Because these are new questions and new algorithms, the evaluation team, EEAC, and PAs recommend conducting sensitivity analysis. For example, in the 'Intent' section of the participant FR battery of questions, don't know and refused responses receive a score of 0.5, which is the midpoint of the intent score. The sensitivity analysis would be useful in deciding whether a different score is more appropriate or whether these participants' responses should be removed from the analysis.

4.1 Participant Market Event Type

Participant Market Event Type questions are included in the NTG batteries to identify whether the high-efficiency equipment installed was due to new construction/major renovation, to replace existing equipment that was still functioning (early replacement), because existing equipment failed (ROF), or because the respondent installed a piece of equipment that was not previously used in their home (new equipment). Responses to these questions are used to identify whether participants should be asked the timing questions in the FR battery.[41]

While the questions do identify rebates that were installed as part of a new construction home, we do not expect many instances where this will occur. This is because the equipment is captured through the New Construction Program and is therefore not eligible for rebates in another program. We include these questions in the algorithm to capture all the possible scenarios so interviews can flow smoothly, and to allow us to measure the extent of any new construction cases that may receive a rebate outside the New Construction program. Cases identified as new construction through this algorithm should be reviewed by the team responsible for evaluating the New Construction program.

The Participant Market Event Type questions were determined by reviewing the existing event type battery used in the prior C&I omnibus SR survey (TXC49), the results of the findings of the 2019 C&I Gas Equipment Early Replacement Study (MA19C11-G-ERCUSTSRV)[42], and event

[41] The event type questions can also be used for other purposes – for example, to give the PAs information to update the percent of units that are early retirement.

[42] Interview findings from the MA C&I Early Replacement Study prepared by the C&I Market Assessment Team (Cadeo and Navigant), February 4, 2020.

type questions identified in the literature review. The recommended market event type questions can be found in Table 2. The algorithm for determining event type can be found in Appendix C.

CAVEAT

Market event type questions do not need to be asked for some types of measures, like insulation, where event type does not impact timing.

Table 2: Market Event Type Questions

Market Event			
ET1	Was the high-efficiency <measure> installed as part of a new construction or major renovation project? (select one)		
	01	Yes	[new construction] SKIP TO Next Section
	02	No	
	88	Don't know	
	99	Refused	
ET2	Did the high-efficiency <measure> you installed replace any existing <equipment> or was it a new type of equipment that you did not have in your home before? (select one)		
	01	Replaced existing equipment	
	02	New equipment	[new equipment] SKIP TO Next Section
	88	Don't know	
	99	Refused	
ET3	[ASK IF ET2=1, DK, R] Which of the following best describes the condition of your old equipment? (read list)		
	01	The old equipment was working with no need of repair	
	02	The old equipment was working with need of minor repair	
	03	The old equipment was working with need of major repair	[ROF]
	04	The old equipment was no longer working	[ROF]
	88	Don't know	
	99	Refused	
ET4	[ASK IF ET3 = 1, 2, DK, and R] Do you think your old equipment would have lasted another two years? (select one)		
	01	Yes	[ER]
	02	No	[ROF]
	88	Don't know	[ROF]
	99	Refused	[ROF]

4.2 PARTICIPANT FREE-RIDERSHIP

The participant FR score consists of two components: an intent score and an influence score. These are followed by consistency checks.

CAVEAT

The wording in the questions shown below are for incentive-based measures. The wording of questions for direct install measures received free as part of an audit will need to be changed to include the phrase "If you had not received the free <measure(s)> during the audit, …".

4.2.1 Intent Score

The intent score incorporates a series of questions to determine whether, in the absence of the program, the respondent:

- would have purchased any type of equipment at the same time, a later date, or not at all (timing)
- would have purchased the same, lesser, or greater quantity of the high-efficiency equipment (quantity)
- would have purchased the same or lesser efficiency equipment (efficiency).

Questions used to determine timing, quantity, and efficiency are contained in Table 3, along with exceptions that evaluators should make depending upon the type of high-efficiency measure purchased.

Table 3: Timing, Quantity, and Efficiency (Intent) Questions

Timing (Evaluator: if NC or ROF, do *not* ask participants the timing questions. New equipment and early replacement should be asked the timing questions.)

T1 Without the <list of all rebates/incentives/support received> from <Sponsor>, how likely is it that you would have installed <u>any</u> <u>type</u> of <equipment> at the same time? Would you say …

01	Not at all likely	FR_timing = 0 (skip to 'influence')
02	Slightly likely	
03	Somewhat likely	
04	Very likely	FR_timing = 1.0 (skip to Q1)
88	Don't know	FR_timing = 0.5 (skip to Q1)
99	Refused	FR_timing = 0.5 (skip to Q1)

T2 Without the <list of all rebates/incentives/support received> from <Sponsor>, when do you think you would have installed the <equipment>? Would you say …

01	Within 6 months of when you did	FR_timing = 1.0
02	Between 6 months and 12 months of when you did	FR_timing = 0.5
03	More than 1 year after when you did	FR_timing = 0 (skip to 'influence')
88	Don't know	FR_timing = 0.5
99	Refused	FR_timing = 0.5

Quantity (Evaluator: Ask only if quantity purchased/received is greater than 1; for example, would not need to ask if a customer purchased one furnace/boiler or if they received a tune-up. For insulation/weatherization type measures ask this series but the wording should be revised to "amount of" instead of "quantity of.")

Q1 Without the <list of all rebates/incentives/support received> from <Sponsor>, how likely is it that you would have installed the exact same [quantity of / amount of] high-efficiency <measure>? Would you say it is …

01	Not at all likely	FR_qty = 0
02	Slightly likely	
03	Somewhat likely	
04	Very likely	FR_qty = 1.0
88	Don't know	FR_qty = 0.5
99	Refused	FR_qty = 0.5

Q2a [non-insulation/weatherization type measures] Without the rebate or assistance from <Sponsor>, how many high-efficiency <measure> would you have installed?

__	Quantity would have installed	FR_qty = (Q2a/installed quantity)
88	Don't know	FR_qty = 0.5
99	Refused	FR_qty = 0.5

Q2b [insulation/weatherization type measures] Without the rebate or assistance from <Sponsor>, what percent of the high-efficiency <measure> would you have installed?

__	Percent would have installed [0-100]	FR_qty = Q2b
88	Don't know	FR_qty = 0.5
99	Refused	FR_qty = 0.5

Efficiency (Evaluator: Ask only if efficiency applies; for tune-up, insulation/weatherization type measures, air sealing type measures, controls, etc. skip to 'influence' questions)

E1 Without the <list of all rebates/incentives/support received> from <Sponsor>, how likely is it that you would have installed the exact same high-efficiency level of <measure>? Would you say …

01	Not at all likely	FR_eff = 0
02	Slightly likely	FR_eff = 0.25
03	Somewhat likely	FR_eff = 0.5
04	Very likely	FR_eff = 1.0
88	Don't know	FR_eff = 0.5
99	Refused	FR_eff = 0.5

The intent score is the minimum of the Timing, Quantity, and Efficiency scores. Note that if one or more of these questions are not asked because they are not applicable, the evaluator should conduct sensitivity analysis to determine what impact only using one or two of these three scores has on the intent score.

4.2.2 Influence Score

The influence score shown in Table 4 asks how influential different program attributes were in the participant's decision to install the high-efficiency measure through the program. The number of program attributes in question X1 will vary depending upon the program's design. The evaluator should review logic models and program theory and conduct staff interviews to inform the list of elements to include. The more typical elements programs used to influence customer decision-making include information; incentives or rebates; interaction with program staff (technical assistance); building audits or assessments; and financing. Depending on the design of the program and what role contractors play in promoting energy-efficient equipment through the program, questions X2 and X3 may get asked. These questions capture the influence the contractor or salesperson had on the decision and help determine how the participant decided on what efficiency of equipment to install. The evaluator should decide whether questions X2 and X3 get asked.

Table 4: Influence Questions

Influence: (The evaluator should review logic models and program theory, and conduct staff interviews to inform the list of elements to include)	
X1	On a scale of 0 to 10 where 0 is 'not at all influential' and 10 is 'very influential,' how influential was < attribute from list below> on your decision to install the high-efficiency <measure>? [NOTE: program features asked about should be tailored to the program being evaluated; list max of 6 factors] a. the program rebate b. the program's marketing materials c. [Evaluation team will need to insert other features of the program design/delivery] __ Rating 66 Not applicable 88 Don't know 99 Refused
X2	[ASK IF PROGRAM THEORY SHOWS CONTRACTOR/SALESPERSON PLAYS A PROMINENT ROLE IN DELIVERING AND PROMOTING THE PROGRAM] Using the same 0 to 10 scale where 0 = 'not at all influential' and 10 is 'very influential,' how influential was the contractor or salesperson on your decision to install the high-efficiency <measure>? __ Rating 66 Not applicable 88 Don't know 99 Refused

X3 [ASK IF CONTRACTOR/SALESPERSON HAS THE HIGHEST INFLUENCE SCORE AND PROGRAM THEORY SHOWS CONTRACTOR/SALESPERSON PLAY A PROMINENT ROLE IN DELIVERING AND PROMOTING THE PROGRAM] Which of the following best describes how you selected the new high-efficiency <measure>? (Select one) [rotate options 1 through 3]

01 I did some research on <measure>s and made my own choice
02 My contractor suggested one <measure> efficiency level, and I agreed – (use Contractor FR results)
03 My contractor suggested various <measure> efficiency levels, and I chose one (average of contractor and customer FR results)
04 Something else (specify)

The influence score is based on the program attribute that is rated highest in question X1. The influence score calculation is shown in Table 5.

Table 5: Calculation of the Influence Score

Rating	Score
Not at all influential – 0	1.0
1	0.9
2	0.8
3	0.7
4	0.6
5	0.5
6	0.4
7	0.3
8	0.2
9	0.1
Very influential – 10	0.0

4.2.3 Participant Free-ridership Score

The Participant FR score is the average of the intent and influence scores (Intent Score + Influence Score / 2).

4.2.3.1 Participant and Contractor Free-ridership Reconciliation

If the participant is asked questions X2 and X3, their response to these questions may impact whether, and how, the contractor FR score is used. If the respondent says the contractor was the most influential factor in their decision (i.e., X2 is rated higher than any program attributes in X1) and, in question X3, says the contractor suggested one measure efficiency level, and the customer agreed, the evaluator should use the contractor FR score instead of the customer's responses to the intent and influence questions. If the respondent in question X3 said the contractor was the most influential program element and the contractor suggested various measure efficiency levels, and the customer chose the measure, the evaluator should average the contractor and customer FR results. See the Participant and Contractor Free-ridership Reconciliation algorithm in Appendix C for details.

4.2.4 Consistency Questions

The FR battery, shown in Table 6, also includes two consistency questions to ensure that responses to the intent and influence questions are consistent. To minimize respondent burden, these questions are only asked in cases where responses to the intent and influence questions are inconsistent. For example, if the respondent says in the intent questions that they would not have purchased any equipment within a year, or they would not have purchased the same quantity and efficiency of equipment, but then say in the influence questions that the program elements asked about (e.g., program rebate, audit, marketing materials) were not influential (a rating of 0, 1, 2, or 3), the consistency questions should be triggered. As another example, if the respondent says in the intent questions that they would have installed the exact same efficiency and quantity of equipment at the same time, but rate one or more of the program elements in the influence questions as highly influential in their decision to install the efficient equipment (a rating of 7, 8, 9, or 10), the consistency questions should be triggered.

While not all jurisdictions like or incorporate consistency checks, we are proposing to include these questions until the battery has been thoroughly tested to ensure the questions capture the necessary elements. These questions could be removed in future studies, assuming they are not found to be heavily relied upon to make adjustments.

Table 6: Consistency Questions

Consistency	
C1	**(CONSISTENCY QUESTION)** Did you make the decision to install the high-efficiency <measure> before or after you learned of the rebate? (select one) 01 Before 02 After 88 Don't know 99 Refused
C2	**(CONSISTENCY QUESTION)** Please describe the influence that the program <list of all rebates/incentives/support received) had on your decision to install the high-efficiency <measure>? [Record verbatim response]

The team conducting the evaluation should examine responses to the consistency questions and use expert judgment to try to resolve the inconsistency and determine if the program should receive any credit. If the inconsistency cannot be resolved, we recommend removing the respondent from the FR estimate. During reporting, the evaluation team should document decisions made and the rationale behind those decisions.

4.3 PARTICIPANT SPILLOVER

In the participant spillover questions, respondents are asked the type and efficiency of equipment purchased, the importance of the program on their decision to purchase the additional efficient equipment, and the likelihood of purchasing the additional efficient equipment had they not also purchased the rebated equipment. The participant spillover questions are heavily based on the prior TXC34 Massachusetts spillover battery.

In the event that Massachusetts is able to conduct the participant FR questions soon (e.g., within six months) after the equipment purchase, the evaluator should consider using a separate survey to estimate participant spillover so that the participants have enough time to make a purchase on their own.

The spillover questions can be found in Table 7.

Table 7: Participant Spillover Questions

Participant Spillover[43]	
S1	Since installing the equipment, have you made other energy-saving purchases or changes that did not receive a rebate through the Mass Save® program? (Select one) 01 Yes 02 No (skip to next section) 88 Don't know (skip to next section)
S2	Did your experience with the Mass Save program influence your decision to take any of these energy-saving actions? (Select one) 01 Yes 02 No (skip to next section) 88 Don't know (skip to next section)
S3	What energy-saving purchases or changes did you make? (Select all that apply) [Note: Make light bulbs and other upstream products available for interviewers and respondents, but these measures will not be counted towards spillover] 01 LEDs/Energy-efficient light bulbs 02 Energy-efficient appliance(s) (specify type) 03 HVAC equipment (specify type) 04 Attic, wall, or basement insulation 05 Duct sealing or duct insulation 06 Air sealing of leaks 07 Smart thermostat 08 Other purchases or changes (specify)

[43] NOTE: To be able to calculate spillover, at a minimum, the survey demographics should collect information on water heater fuel, clothes dryer fuel, primary heating fuel, and type of cooling equipment in order to determine what prescriptive savings to apply.

[ASK S3b to S6 FOR EACH MEASURE OTHER THAN LIGHTING[44] SELECTED IN S3]

S3b [if S3 = 2 or 3 or 8] Can you describe the <equipment from S3> work in more detail? If applicable, include the quantity installed, the type of equipment, and the efficiency levels installed. (For AC probe on SEER rating, for furnaces or boilers probe on AFUE, for heat pumps probe on Heating season performance factor [HSPF] or SEER)

S3c How did you know the equipment was energy efficient and would have qualified for a rebate through the program?

[Record verbatim response]

S3a [If S3=2 8and ENERGY STAR not mentioned in S3b] Was the <equipment from S3> ENERGY STAR-labeled? (Select one)

01	Yes	
02	No	
88	Don't know	

S4 How important was your experience with the Mass Save program on your decision to install <equipment from S3> that did not receive a rebate from <Sponsor>? Was it ...

01	Not at all important	score = 0
02	Slightly important	score = 3
03	Somewhat important	score = 6
04	Very important	score = 10
66	Not applicable	
88	Don't know	
99	Refused	

S5 How likely is it that you would still have purchased/installed energy-efficient <equipment from <S3> if you had *not* already received a rebate for the energy-efficient <measure>? Would you have been ... to have installed the energy-efficient <equipment from S3>?

01	Not at all likely	score = 10
02	Slightly likely	score = 6
03	Somewhat likely	score = 3
04	Very likely	score = 0
66	Not applicable	
88	Don't know	
99	Refused	

S6 **(Consistency question)** Why did you not submit an application for a rebate through the Mass Save program for this additional equipment?

[record verbatim response]

[Note: for respondents who indicate that the equipment did not qualify, no spillover should be attributed to the program]

[44] LEDS are assumed to be counted as part of the upstream program. If the upstream program is ended, then they should be counted towards SO.

Based on responses to S3b, evaluators should align efficiency specifications with TRM specifications to determine measure savings. Secondary sources should be used to estimate savings for non-like measures if they are not available in the MA TRM.

Where spillover is found, a follow-up telephone or web survey may be used to allow participants to submit pictures. The pictures could help determine if the equipment qualifies for spillover. They could also help detect savings that may be attributed to the program. Incentives may need to be considered if this approach is used.

The spillover score is calculated as follows. If the score is 5 or lower, spillover is set to 0. If the score is higher than 5, the spillover savings is the appropriate TRM (or secondary sources) savings.

Participant Spillover score = S4 + (10-S5) / 2

CAVEAT

To avoid double-counting, evaluation teams will need to review spillover mentioned by participants and contractors at the measure-level. If spillover is calculated for the same measures by both participants and contractors, the evaluator should subtract any savings reported by the participants from the contractor spillover estimate. We base this conservative approach on the assumption that participating design professionals or vendors were reporting the same information as the participant's spillover project. This is consistent with how the C&I SR NTG calculation works.

Where the evaluator has evidence that spillover may be high, this may warrant a separate study.

4.4 Contractor Free-ridership

For measures or programs where the program theory shows contractors play a prominent role in delivering and promoting the program, such as HVAC equipment in residential retrofit, contractors will be asked about the influence of the program on their sales. This approach was utilized in the recent Residential HVAC Net-to-Gross and Market Effects Study (TXC34).

The contractor FR battery starts by confirming the contractor's sales of the program-qualifying equipment. The contractor is then asked how their sales would have changed if the program rebates had not been available. The questions can be found in Table 8.

Table 8: Contractor Free-ridership Questions

Contractor Free-ridership	
CF0	For the next set of questions, we are going to ask you to think about your company's experiences with Mass Save rebates. Our records show that in <year> your company sold <QTY> residential high-efficiency <measure> that received rebates from Mass Save. Does this sound right? 01 Yes [program sales = QTY] skip to CF1 02 No 88 Don't know [program sales = QTY] skip to CF1
CF0a	About how many residential <measures> that received rebates from Mass Save did your company sell in <year>? ____ [0 TO 1000] [program sales = CF0a] 8888 Don't know [skip to next section] 9999 Refused [skip to next section]
CF1	If Mass Save had not offered any rebates in <year>, about what percentage of the <QTY> high-efficiency <measure> that received rebates would you have sold that year? [Interviewer note: If the contractor only sold 1, then the response should be 0% if said would not have sold any, or 100% if said would have sold the same amount] ___ [0 TO 100] 888 Don't know [skip to next section] 999 Refused [skip to next section]
CF2	[IF CF1 < = 100] To confirm, you are estimating that in <year>, your company likely would still have installed <QTY * (CF1 / 100)> <measure> of the <QTY> you sold if Mass Save had not offered the rebates. Is that **roughly** correct? 01 Yes [skip to next section] free-ridership = CF1 02 No 88 Don't know [skip to next section] 99 Refused [skip to next section]
CF3	[IF CF2 = 2] Could you provide your best estimate of about how many of the <QTY> <measure> your company would still have sold in <year> if Mass Save had not offered the rebates? ___ [0 TO 1000] 8888 (Don't know) **Free-ridership = CF3/program sales**

Responses to these questions will be used to calculate a free-ridership score that will replace the participant FR score or may be averaged with the participant score. Where participants rated the contractor highly influential in doing the project and indicated they used the contractor's equipment recommendations, the contractor FR score will be used in place of the participant FR score. Similarly, when the contractor was highly influential in the participant's decision but the customer ultimately made the equipment decision themselves with input from the contractor, the

participant and contractor FR scores will be averaged. This allows the program to factor in the impact the contractors have had on the selection of the program qualifying-equipment.

This study has been designed to run concurrently with the participant FR study. If there is ample time, the evaluation team could run the contractor FR study upon completion of the participant FR study. In doing so, only measures where customers indicated the contractor was highly influential in the decision would need to be included in the contractor FR study.

4.5 Participating Contractor Spillover

The contractor spillover questions are asked of participating contractors.[45] These questions are used to assess the influence of the program on the contractor's recommendations and the influence of the program and other factors on the number of sales of program-eligible equipment outside the program, after capturing sales information. This approach is similar to that used in the prior TXC34 Massachusetts study. The questions can be found in Table 9.

Table 9: Participating Contractor Spillover Questions

NS1	Think of all your sales of residential <measure>, including standard-efficiency and high-efficiency and rebated and non-rebated units. About how many did your company sell **in total** in <year>?		
	____	Total installations	
	8888	Don't know	[skip to next section]
	9999	Refused	[skip to next section]
NS2	About how many of the sales of residential <measure> in <year> were high-efficiency, including rebated and non-rebated units?		
	___	Number of installs	
	8888	Don't know	[skip to next section]
	9999	Refused	[skip to next section]
NS3	Did all the high-efficiency <measure> that your company sold in <year> receive a Mass Save rebate of some kind?		
	01	Yes	[skip to next section]
	02	No	
	88	Don't know	[skip to next section]
	99	Refused	[skip to next section]

[45] The evaluation team could also conduct surveys with non-participating contractors to assess the influence of programs on contractors' sales of efficient equipment. Depending upon the program theory or changes in the market, these questions may be considered for future studies.

NS4 About what percentage of the high-efficiency <measure> that your company sold in <year> did not receive a Mass Save rebate even though they would have qualified for one?

___ Percent [0 TO 100]
888 Don't know
999 Refused

NS5 [ASK IF NS2 AND NS4 < 888] To confirm, that means about <(NS2) * (NS4 / 100)> of the <NS2> high-efficiency <measure> your company sold in <year> did not receive **a Mass Save rebate** even though they would have qualified for one. Is that **roughly** correct?

01 Yes
02 No
88 Don't know
99 Refused

NS6 [ASK IF NS5 = 2, 88, OR 99] About how many of the sales of residential high-efficiency <measure> in <year> did not receive a rebate even though they qualified?

____ [OPEN END NUMERIC]
8888 Don't know
9999 Refused

[NOTE: IF NONE SOLD OUTSIDE PROGRAM ((NS6 = 0, 8888, OR 9999) OR (NS3 = 1) OR (NS4 = 0)), SKIP TO NEXT MEASURE/SECTION]

NS7 Next, I'd like you to think about the energy-efficient equipment you sold that did <u>not</u> receive a rebate. Using a 0 to 10 scale where 0 is 'not at all influential' and 10 is 'very influential,' how influential were the Mass Save rebates on your sales of high-efficiency <measure> in <year>?

__ 0 to 10 Rating
88 Don't know
99 Refused

NS8 Using that same scale (where 0 is 'not at all influential' and 10 is 'very influential'), how influential was the program support such as marketing, advertising, education and training on your company's sales of these high-efficiency <measures>?

__ 0 to 10 Rating
88 Don't know
99 Refused

The influence score is based on the maximum rating in NS7 and NS8. The non-participant spillover score is then calculated as:

(Non-program sales x (Influence score / 10)) / Program Sales

CAVEAT

As mentioned in the Participant SO section above, evaluation teams will need to review any spillover mentioned by participants and contractors at the measure-level to avoid any double-counting of savings. Any spillover calculated for the same measures that is reported by both participants and contractors will need to be reviewed and handled appropriately. The evaluator should subtract any savings reported by the participants from the contractor spillover estimate to conservatively avoid any double counting.

Appendix A Addendum to Literature Review: SR NTG Methods and Algorithms used in Leading States

This appendix provides more detail on the methods and algorithms used in other states. Note that due to the amount of detail contained in these documents, much of this information is quoted verbatim from the methods documents. **Unless the wording clearly indicates otherwise, readers should assume that the text in each section is either a direct quote or a paraphrase from the source documents.**

A.1 Energy Trust Free-Ridership Methodology[46]

The discussion below summarizes the historic NTG measurement for Energy Trust of Oregon. Per Phil Degens from Energy Trust of Oregon, NTG will not apply after December 31, 2019. They still plan to look at program influence, but the focus will be for program redesign or planning purposes. For purposes of our review of residential NTG batteries and algorithms, we felt the historic information below would be useful.

Energy Trust staff has developed a method for calculating free-ridership that they believe is simple, transparent, and unbiased. One goal in developing this method was the ability to apply it to all programs and their markets. An additional goal was the ability to obtain the SR results through a reduced set of survey questions. These questions are now part of Energy Trust's Fast Feedback survey, a short phone survey administered shortly after program participation on an ongoing basis.

In piloting this survey, Energy Trust staff reasoned that the timing of the survey, as well as its brevity (the survey averaged five minutes in 2012), would increase participant response rates. The survey is administered about one month after participants receive their incentive checks.

According to the methodology, a project's free-ridership score consists of two elements: a stated intent/project change score and an influence score. For residential programs, these scores are calculated based on respondents' answers to two questions.

The remaining sections are largely quoted from the Energy Trust's Free-Ridership Methods document.

[46] Phil Degens, Sarah Castor, and Erika Kociolek, Updated August 7, 2013 https://www.energytrust.org/wp-content/uploads/2016/12/Energy_Trust_Free_Ridership_Methods.pdf

A.1.1 Stated Intent in the Absence of the Program / Project Change

The stated intent/project change score is based on the respondent's answer to a question about how the project would have changed in the absence of the program: "Which of the following statements describe the actions you would have taken if Energy Trust incentives and information were not available?"

Possible responses to this question vary slightly by program and measure, and fall into one of three categories representing different levels of change:

1. The participant would have done exactly the same thing (no change)
2. The project would have changed but retained some energy-efficiency features (some change)
3. The project would have made other changes with no significant energy-efficiency component (significant change)

Energy Trust staff assigned a number between 0 and 1 to each category, where 0 indicates no free-ridership and 1 indicates the project was a full free rider (see Table 10). Changes that might have retained some of the energy-efficiency features of the project were scored at the midpoint as no reliable information on the efficiency level was available.

Table 10: Free Rider Scoring of Stated Intent / Project Change

Response	FR Score
No change in project	1
Some change	0.5
Significant change	0
Don't know*	-

A.1.2 Program Influence

The influence score is based on respondents' answers to questions about the influence of various elements, including Energy Trust incentives, program representatives, contractors or salespersons, studies, and program- or measure-specific elements.

Respondents rate each element on a five-point scale, where 1 is *not at all influential* and 5 is *extremely influential*. The answer choices are given a score between 0 (element's influence was a 5, extremely influential) and 1 (element's influence was a 1, not at all influential) – see Table 11. The score for the most influential element is the influence score.

Table 11: Free Rider Scoring of Program Influence

Response	Score
Extremely influential (5)	0
4	0.25
3	0.5
2	0.75
Not at all influential (1)	1
Don't know*	-

A.1.3 Free Rider Rate

With the outcomes of being influenced or not being influenced by the program having equal probabilities, the free rider rate associated with each outcome are additive:

Free Rider Rate = probability(Program had influence) +
probability(Program had no influence)

Free Rider Rate = 50%(Score associated with stated intent/project change) +*
50%(Score associated with program influence)*

Free Rider Rate = Free rider rate for stated intent/project change outcome +
Free rider rate for program influence outcome

The rate is obtained by multiplying the free rider score by 50%. For the stated intent/project change question, 50% represents the probability that the program had no influence. For the program influence question, 50% represents the probability that the program had influence.

A.2 Evaluation Framework for Pennsylvania Act 129 Phase III Energy Efficiency and Conservation Programs[47]

In Pennsylvania, the Evaluation Framework follows Pennsylvania Act 129 in requiring NTG measurement for the purposes of program modifications, program planning, and determining program cost-effectiveness, but not for compliance with targets.

The PA Framework states that when conducting NTG research, the NTG methods should be consistent across time and electric distribution companies (EDCs). The SWE has determined that EDCs should use SRA for assessing free-ridership and spillover for downstream programs and the PA Framework provides descriptions of common methods for doing those assessments. It notes that these approaches must be used for the specific programs they apply to, though they may be used in combination with other methods.

In addition to using the SRA in the common approach, the PA Framework also discusses the use of econometric approaches and market effects studies. But the SWE added that because of the cost of econometric approaches and market effects studies and the inability to disaggregate the effects for FR and SO, they determined that primarily survey methods should be used for measuring FR and SO for downstream programs, and that other methods can be used for upstream programs or to provide information on market effects.

The SWE considers the research approaches described in the UMP, as well as those used in Massachusetts and by Energy Trust of Oregon to constitute some of the best practices for free-ridership and spillover estimation.

[47] Evaluation Framework for Pennsylvania Act 129 Phase III Energy Efficiency and Conservation Programs, Prepared by NMR Group, Inc., Ecometric Consulting, LLC, Demand Side Analytics, LLC, August 25, 2016. http://ilsagfiles.org/SAG_files/Technical_Reference_Manual/Version_7/Final_9-28-18/IL-TRM_Effective_010119_v7.0_Vol_4_X-Cutting_Measures_and_Attach_092818_Final.pdf

The SWE determined that EDCs should use standard sampling techniques, data collection approaches, survey questions, survey instruments, and analysis methodology for free-ridership assessment. One common approach applies to a broad range of incentive-based programs, and the other is specific to appliance recycling programs.

The remainder of this section is largely quoted from the PA Framework document.

A.2.1 Free-Ridership for Incentive-Based Programs

The common approach for incentive-based programs is similar to that chosen by Energy Trust, which uses a short battery of questions. The core NTG is calculated as 1 - FR + SO + ME. It provides an SRA for the common approach to measure FR for downstream programs, specifying survey questions to assess participant intention (no-program)[48] and program influence[49]. These two components are additive and equally weighted. Each component provides a possible score of 0 to 50. When added, the resulting score is interpreted as a free-ridership percentage. The PA Framework notes that these two components were chosen as a way to balance potentially different and opposing biases -- the intention component typically indicates higher free-ridership than the *influence* component. The logic is that combining the two components decreases the potential biases.

This common method can be adapted for different types of variations of programs or measure types. The PA Framework discusses alternate wording for programs that do not fit the standard retrofit incentive model (e.g., direct install, financing).

Intention. As described in the PA Framework, *intention* is assessed through a few brief questions used to determine how the upgrade or equipment replacement likely would have differed if the respondent had not received the program assistance. The initial question asks the respondent to identify, from a limited set of options, what most likely would have occurred without the program assistance. Note that *program assistance* often includes more than just the incentive or rebate – it may also include audits, technical assistance, and the like.

Examples of response options (typically four or five, and preferably no more than six) capture the following general outcomes:

- Would have canceled or postponed the project, upgrade, purchase, etc., beyond the current program cycle (typically at least one year).
- Would have done something that would have produced savings, but not as much as those achieved through the upgrade or equipment replacement as implemented.
- Would have done the upgrade or equipment replacement as implemented.
- Don't know

The first outcome (canceled or postponed beyond the program cycle) indicates zero free-ridership and thus results in a score of 0. The second option indicates some free-ridership, but not total free-ridership (a score ranging from 12.5 to 37.5 for the *intention* component). The level of free-

[48] Intention (no program) reflects the customer's intention to carry out the energy-efficient project without program funds.
[49] Influence reflects the influence of the program in the customer's decision to carry out the energy-efficient project.

ridership depends on two factors: (1) the level of savings that the respondent would have achieved without the program's assistance and (2) in the case of non-residential programs, whether the respondent's business or organization would have paid the entire cost of the equipment replacement or upgrade without the program assistance. The third outcome (done project as implemented) indicates total free-ridership (a score of 50 for the *intention* component).

The PA Framework contains detailed instructions for assessing intention for different program options. The assessment of *intention* for residential programs is similar to that for non-residential programs. However, the response option "reduced the size, scope, or efficiency of the project" is not likely to be as meaningful to a residential respondent as to a non-residential one, nor is a residential respondent expected to be able to estimate whether the reduction would be small, moderate, or large. Thus, evaluators should attempt to provide a list of meaningful counterfactual options.

Table 12 shows examples of counterfactual response options used with three types of residential measures: appliances, air or duct sealing or insulation, and windows. As this shows, the goal is to cover the range of likely alternatives to carrying out the incented upgrade, with intention scores that reflect the degree of free-ridership. Reporting an alternative that likely would have produced no energy savings results in a score of 0; reporting something that likely would have produced some energy savings, but lower savings than the incented upgrade or purchase results in an intermediate score of .25; and reporting the same outcome as the incented upgrade or purchase results in a score of 0.5.

Table 12: Example Counterfactual Response Options for Various Residential Measure Types

Counterfactual Responses		Intention Score
Appliance	Cancel/postpone purchase	0
	Repair old appliance	0
	Buy used appliance	0
	Purchase less expensive appliance	0.25
	Purchase less energy-efficient appliance	0.25
	Purchase same appliance without the rebate	0.5
	Don't know	0.25
Air/Duct Sealing, Insulation	Cancel/postpone	0
	Do by self (if program incents only contractor-installation)	0.25
	Reduce amount of sealing/insulation	0.25
	Have the same level of sealing/insulation done without the rebate	0.5
	Don't know	0.25
Windows	Cancel/postpone purchase	0
	Replace fewer windows	0.25
	Purchase less expensive windows	0.25
	Purchase less energy-efficient windows	0.25
	Do same window replacement without the rebate	0.5
	Don't know	0.25

Influence. The PA Framework also details the questions to use for measuring the influence of different program components. Program influence may be assessed by asking the respondent how much influence – from 1 (no influence) to 5 (great influence) – various program elements had on the decision to implement the energy-efficiency project. The number of elements included varies depending on program design. Logic models, program theory, and staff interviews typically inform the list of elements to include. Among the more typical elements programs used to influence customer decision making include information; incentives or rebates; interaction with program staff (technical assistance); interaction with program proxies, such as members of a trade ally network; building audits or assessments; and financing.

The program's influence score is equal to the maximum influence rating for any program element. The rationale is that if any given program element had a great influence on the respondent's decision, then the program itself had a great influence, even if other elements had less influence. The influence score has an inverse relationship with NTG – the greater the influence score, the lower the free-ridership. The influence scores for each rating are shown in Table 13.

Table 13: Response Options and Associated Score for Influence

Response	Score
1 – Not at all influential	50
2	37.5
3	25
4	12.5
5 – Extremely influential	0
Don't know	25

The PA Framework recommends against consistency checks. The PA Framework acknowledges that consistency checks are frequently used in social and epidemiological research but feels it may not result in more accurate data. The PA Framework states that the assumption that the inconsistency can be resolved accurately may be unfounded. That assumption is based on the belief that the questioner can accurately and reliably determine which of two inconsistent responses is the correct one. A respondent confronted with inconsistent responses may seek to resolve the consistency, but that does not mean that the final response will be accurate. Instead, the response may be influenced by *self-enhancement* motivation.

The PA Framework also notes that if respondents are confronted with inconsistent responses, this may impact later responses to the survey and lengthen the survey. Further, the need to assess which response is correct brings more evaluator subjectivity into the assessment. Therefore, the PA Framework recommends against consistency checks.

Total free-ridership is the sum of the intention and influence components, multiplied by .01 to convert it into a proportion.

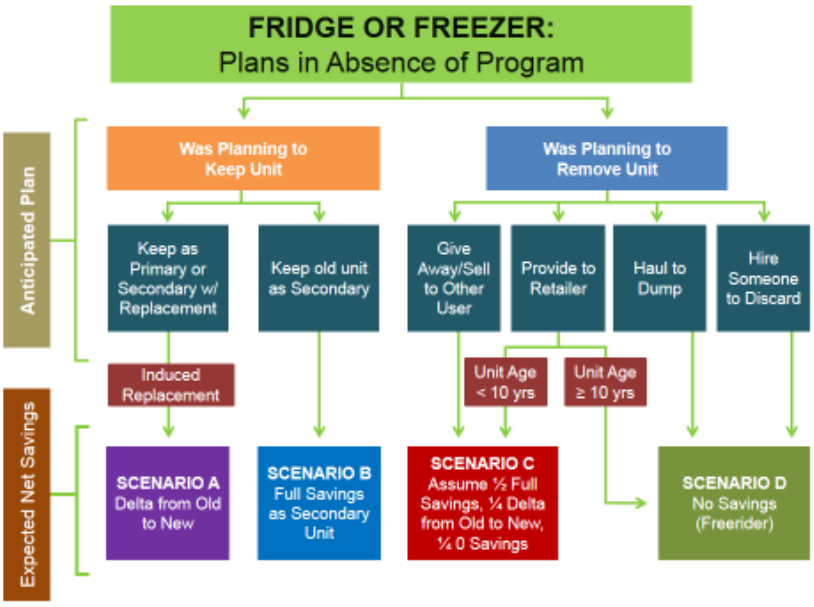
A.2.2 Free-Ridership for Appliance Recycling Programs

The approach for the Appliance Recycling programs is based on the approach described by the UMP because the program theory and logic for appliance retirement differs significantly from standard downstream incentive programs. This approach is also used for appliance replacement with early retirement programs.

Per the PA Framework, net savings for the appliance retirement program is based on the participants' self-reports of what they would have done absent the program. Savings are attributed based on four scenarios: (1) they would have kept the unit in the absence of the program but instead, as a result of the program, replaced it with a more efficient one (savings equals delta energy usage from old to new unit); (2) they would have kept the unit in the absence of the program but instead, as a result of the program, recycled it and did not replace it (savings equals energy usage of old unit); (3) in the absence of the program, they would have put the unit back into usage elsewhere, sold or given the unit away to another user, or sold or given away a unit that was less than ten years old to a retailer (savings equals a mix of full savings, delta old to new, and no savings); or (4) in the absence of the program, they would have taken the unit out of usage, sold or given a unit at least ten years old to a retailer, hauled it to the dump, or hired someone to discard it (free rider – no savings).

Figure 1 shows how the net savings are derived for the appliance retirement program using the 2013 UMP chapter. Note that this approach is slightly different in the 2017 revised UMP chapter.

Figure 1: Pennsylvania Protocol for Appliance Recycling



The PA Framework also notes that general population surveys can be useful to offset some of the potential bias in estimating the true proportion of the population that would have recycled their unit in absence of the program. While a non-participant survey is recommended, it is not required given budget and time considerations.

A.2.3 Spillover

The SWE also determined that EDCs should use standard techniques, instruments, and methods for spillover assessment where possible. Estimation of non-participant spillover is desirable through a general population survey or a survey of trade allies; however, it is not required. The PA Framework describes the general form of questions to use and rules for calculating spillover scores from responses to questions.

A.2.3.1 Participant Spillover

The common approach for participant spillover assesses the number and description of non-incented energy-efficiency measures implemented since program participation, an estimate of energy savings associated with those energy-efficiency measures, and the program's influence on the participant's decision to implement the identified measures.

Identification of Non-Rebated Residential Measures: The survey should assess the purchase and installation of any energy-efficient measures, whether eligible for program rebates, in the TRM but not eligible, or not in the TRM. The survey asks participants a series of questions similar to the following to determine whether they installed any additional energy-efficient measures without receiving a rebate:

- You received a rebate for installing [list of rebated measures]. Since participating in the program, have you installed any additional [list of rebated measures] for which you did not receive a rebate?
    - [IF YES:] How many/how much have you installed?
- Since participating in the program, have you installed any other energy-efficient products or equipment, or made any energy-efficiency improvements for which you did NOT receive a program rebate?
    - [IF YES:] What type of other energy-efficient improvements, products, or equipment did you install? [Record description of each additional installed measure]
    - [FOR EACH MEASURE:] How many/how much did you install?

Assessment of Program Influence on Residential Measures: The survey should also ask respondents about the level of influence the prior program participation had on their decision to install the additional measures. The survey may apply a single influence assessment to all measures, under the assumption that residential respondents are not likely to report different levels of program influence for different measures. At the evaluator's discretion, the survey may assess influence for each measure identified.

The SWE recommends that the influence question identify various ways in which the program participation might have influenced the decision to install additional measures. For example, evaluators may consider a question similar to the following: on a 1 to 5 scale, where 1 means *not at all influential* and 5 means *extremely influential*, how influential were each of the following on your decision to [vary wording as appropriate] install the additional equipment/product(s)/improvement(s)?

- Information about energy savings from utility marketing, program representatives, retailers, or contractors
- Your satisfaction with the equipment for which you had received a rebate
- Your installation of [rebated measure(s)] made you want to do more to save energy

Program influence is assessed as the maximum influence rating given to the program elements. The maximum influence rating is assigned a score that determines what proportion of the relevant measures' savings is attributed to the program:

- A rating of 4 or 5 = 1.0 (full savings attributed to the program).
- A rating of 3 = 0.5 (half of the savings attributed to the program).
- A rating of 1 or 2 = 0 (no savings attributed to the program).

Assessment of Energy Savings for Residential Spillover: Where applicable, the savings for each additional measure installed should be calculated per the TRM for a rebated measure installed through the program. For partially-deemed measures, a working group of the Program Evaluation Group (PEG) will develop conservative working assumptions for any required inputs (e.g., square footage of home, R-value improvement, replaced wattage). As an alternative, the PEG working group may identify average verified savings for such measures.

Evaluators first calculate spillover savings for each spillover measure reported as the product of the measure savings, number of units, and influence score:

$$Measure\ SO = Measure\ Savings * Number\ of\ Units * Program\ Influence$$

For each of the above categories, the evaluators will then make the following calculations:

- Total the savings associated with each program participant, to give the overall participant SO savings.

$$Participant\ SO = \Sigma Measure\ SO$$

- Multiply the mean participant SO savings for the participant sample by the total number of participants to yield an estimated total participant SO savings for the program.

$$\Sigma Participant\ SO\ (population) = \Sigma Participant\ SO\ (sample) Sample\ n \text{ x Population } N$$

- Divide that total savings by the total program savings to yield a participant spillover percentage:

$$\%\ Participant\ SO = \Sigma Participant\ SO\ (population) Program\ Savings \text{ x } 100$$

A.2.4 Non-Participant Spillover

Non-participant spillover may be assessed either through a general population (non-participant) survey or through a survey of trade allies. If a general population survey is selected, it should assess the following, for each survey respondent:

- The number and description of non-incented energy-efficiency measures taken in the program period.
- An estimate of energy savings associated with those energy-efficiency measures.
- The program's influence on the participant's decision to take the identified measures, assessed with a rating scale and converted to a proportion, with possible scores of 0, .5, and 1.

If an evaluator chooses to assess non-participant spillover through trade ally surveys, the survey should assess the following, for each sampled respondent:

- The number of program-qualified measures sold or installed within the specified sector, in the specified utility's service territory, in the specified program year.
- The percentage of such installations that received rebates from the specified program.
- The trade ally's estimate of the proportion of their sales or installations of non-rebated measures that went to prior program participants.
- The trade ally's judgment of the specified program's influence on sales of the common program-qualified but not rebated measures, assessed with a rating scale and converted to a proportion, with a minimum score of 0 and a maximum score of 1.

The survey should estimate total sales of all program-qualified measures by asking trade allies to report sales of their most commonly sold program-qualifying measures and determining what proportion of their total sales of high-efficiency products those measures made up (details in Section, below).Trade ally survey questions should ask about sales within a specific sector (residential or non-residential). If an evaluation plan calls for a single trade ally survey in a given sector to provide SO figures across multiple programs within that sector, that survey should be worded to ensure that the trade ally understands that responses should refer to the multiple programs.

Identification of Non-rebated Measures: The trade ally surveys will ask about sales or installations of the program's most common qualified measures. To reduce burden on respondents, the recommended common method takes the following approach.

First, evaluators should identify each sampled *trade ally's* most commonly rebated measures as well as other commonly rebated program measures of the type pertinent to the trade ally. The survey should assess the number of non-rebated units sold of each of the respondent's most commonly rebated measures within the territory of the EDC in question. To prevent undue burden, the survey should restrict the number of measures investigated to no more than four. For each of those measures, the survey should ask respondents questions similar to the following:

- During the program year, how many [measure] did you sell/install within the service territory of [EDC]?

- Approximately what percentage of your [measure] installations in [EDC] service territory received rebates through the program?

By subtraction, the response to the second question provides the percentage of non-rebated units, of a specific type, sold/installed.

For each of the respondent's most commonly sold program-rebated measures, the number of non-rebated units will be estimated as total number of units sold/installed multiplied by the non-rebated percentage.

The PA Framework states that it is impractical for the survey to attempt to estimate the number of units of *all* program-qualified measures that a respondent sold. This means that the above procedure will underestimate spillover. As a way of providing some information on the possible degree to which spillover is underestimated, the PA Frameworks states that the survey should ask respondents a question like: "Thinking about those types of products together, what percentage do they make up of your total dollar sales of high-efficiency products?" The purpose of this question is not to inform a precise and reliable estimate of additional spillover, but rather to provide information on the possible degree to which spillover is underestimated.

Assessment of Program Influence: For each of the identified measures, the survey will ask respondents about the level of influence the program had on their sales/installations of non-rebated program-qualified measures, using a question similar to the following:

- Using a 1 to 5 likelihood scale, where 1 is *not at all influential* and 5 is *extremely influential*, how influential was the program on your sales of non-rebated high-efficiency products of that type to your customers?

For each measure identified, the maximum influence rating is assigned a score that determines what proportion of the measure's savings is attributed to the program:

- A rating of 4 or 5 = 1.0 (full savings attributed to the program).
- A rating of 3 = 0.5 (half of the savings attributed to the program).
- A rating of 1 or 2 = 0 (no savings attributed to the program).

Assessment of Energy Savings: The savings for each additional measure installed is calculated per the TRM for a rebated measure installed through the program. For partially deemed measures, a working group of the PEG will develop conservative working assumptions for any required inputs (e.g., square footage of home, R-value improvement, replaced wattage). As an alternative, the PEG working group may identify average verified savings for such measures.

Calculation of Trade-Ally-Reported Spillover (SO): For each surveyed trade ally, the total SO of each reported measure (i.e., the commonly rebated measures) will be calculated as follows:

$$Reported\ Measure\ SO = Measure\ Savings * Number\ of\ Units * Program\ Influence$$

The SO from each measure will be summed for each surveyed trade ally to calculate the total SO for that trade ally. Total trade-ally-reported SO for a program can be estimated one of two ways:

- Calculate the mean total SO per trade ally and multiply it by the total number of trade allies, if known, to estimate total SO for the program.
- Calculate the mean SO percentage for each sampled trade ally as the trade ally's total SO divided by the trade ally's total program savings; calculate the mean SO percentage across sampled trade allies (weighted by trade ally size; see below) and multiply that mean SO percentage by the total program savings (from the program database) to estimate total SO for the program.

In either case, the mean total SO or mean SO percentage for trade ally-reported measures should be weighted by trade ally size using total program sales of non-rebated high-efficiency equipment (if available) or by a reasonable proxy, such as total program incentives. The means also should be weighted by trade ally type (e.g., lighting or non-lighting).

Total trade-ally-reported SO can be divided by the total program savings to yield a total SO percentage, as follows:

$$\%\ Total\ Trade\ Ally\ (TA) Reported\ SO = \Sigma Total\ TA\ Reported\ SO\ Across\ all\ Program\ TAs Program\ Savings$$

The evaluators should calculate and report the weighted mean percentage of total sales of high-efficiency equipment that the reported SO measures constitute. The percentage should be weighted by total sales of high-efficiency equipment (if available) or by a reasonable proxy, such as total program incentives. The purpose is to provide a *best available* indication of the degree to which spillover may be undercounted.

Total and Non-participant Spillover

The PA Framework notes that the above approach theoretically yields (but underestimates) total SO because it does not differentiate between sales of non-rebated measures to program participants and non-participants. If responses to the trade ally survey indicate that the trade-ally-identified commonly sold program-rebated measures comprise a large percentage (e.g., 90% or more) of all high-efficiency equipment sold, then evaluators should attempt to determine what percentage of the total trade-ally-identified SO is from non-participants by subtracting the total participant SO for that sector from the total trade-ally-reported SO, as follows:

$$\Sigma Nonparticipant\ SO = \Sigma Total\ TA\ Reported\ SO - \Sigma Participant\ SO$$

That total, divided by the total program savings, yields a non-participant SO percentage, as follows:

$$\%\ Non\text{-}participant\ SO = \Sigma Non\text{-}participant\ SO Program\ Savings$$

If the trade-ally-identified commonly sold program-rebated measures do not comprise a large percentage (e.g., 90% or more) of all high-efficiency equipment sold, then subtracting participant SO will likely not yield an accurate estimate of non-participant SO. In that case, evaluators should report the total trade-ally-reported SO and participant SO.

A.3 ILLINOIS TECHNICAL REFERENCE MANUAL[50]

The Illinois Commerce Commission directed the evaluation teams to compile and formalize consistent NTG methods for use in Illinois EM&V work. The Commission's directives were twofold: (1) assess NTG methodologies and survey instruments that have been used to evaluate energy-efficiency programs, and (2) compile the most justifiable and well-vetted methodologies in an attachment to the updated IL TRM. The Commission noted that the Illinois NTG Methods should be flexible and adaptable to multiple program designs and budgets. It also noted the Methods should be tailored to appropriately assess the specifics of each of the PA's energy-efficiency programs. The statewide NTG methodology document covers the majority of residential and non-residential programs offered in Illinois. If the NTG protocol is not appropriate, instructions are included for diverging from the Illinois NTG Methods. The methodology is updated annually.

Virtually all Illinois based evaluations of residential programs use a survey-based approach for programs where primary data is used to determine net savings. (The main exception is for behavioral programs, which use statistical analysis based on a randomized control trial program design.) Survey-based approaches obtain data from program participants and non-participants using a structured data collection instrument implemented via phone, in person, or online. At times, evaluators create and use an unstructured depth-interview guide to collect information about attribution, and this provides both contextual data and quantitative data about a given project.

The remaining sections are largely quoted from the IL TRM.

A.3.1 Overview of Residential Survey Based NTG Approaches – IL TRM

The IL TRM includes specific protocols for specific types of residential programs. The protocols include Appliance Recycling, Upstream Lighting, Prescriptive Rebate (with no audit), Single-Family Home Energy Audit, Multifamily, Energy Savings Kits and Elementary Education, and Residential New Construction. For income-eligible programs, the NTG value has been deemed at 1.0, unless the Illinois Stakeholder Advisory Group (SAG) or Income Qualified Advisory Committees agree that there is value in performing the NTG research.

The protocols also provide guidance on placement of free-ridership questions; these questions should be asked near the beginning of a participant survey before asking satisfaction questions to prevent participants from confusing free-ridership questions with the satisfaction questions, which could influence free-ridership scores.

The cross-cutting residential protocol formulates the core NTG as 1 – free-ridership (FR) + participant spillover (PSO) + non-participant spillover (NPSO) and provides specific questions and scoring algorithms for measuring each. The cross-cutting protocol provides detail on measuring PSO and NPSO, but it defers to the specific program protocols for measuring FR.

[50] Illinois Statewide Technical Reference Manual for Energy Efficiency, Version 8.0. Volume 4: Cross-Cutting Measures and Attachments. October 17, 2019. https://s3.amazonaws.com/ilsag/IL-TRM_Effective_01-01-20_v8.0_Vol_4_X-Cutting_Measures_and_Attach_10-17-19_Final.pdf

The TRM specifies that the sample for estimating free-ridership should consist of a random sample of current program participants or up to one year of previous program participants. Spillover should be measured within the last 12 months from the survey date.

A summary of the specific protocols for downstream programs follows. The algorithms for PSO and NPSO can be found after this discussion.

A.3.1.1 The Prescriptive Rebate (with No Audit) protocol

This protocol provides basic and enhanced methods with specific questions and scoring algorithms for measuring FR and PSO. Questions assess program influence and no-program components[51] as well as consistency check questions on the program's influence to resolve possible conflicting responses. The basic method measures FR using a customer SRA. The enhanced method provides a protocol to triangulate and develop a weighted combination of FR estimates from two sources: the basic method and a trade ally survey. Trade ally surveys may also be used to assess NPSO. The free-ridership battery in the basic method is brief to avoid undue survey burden.

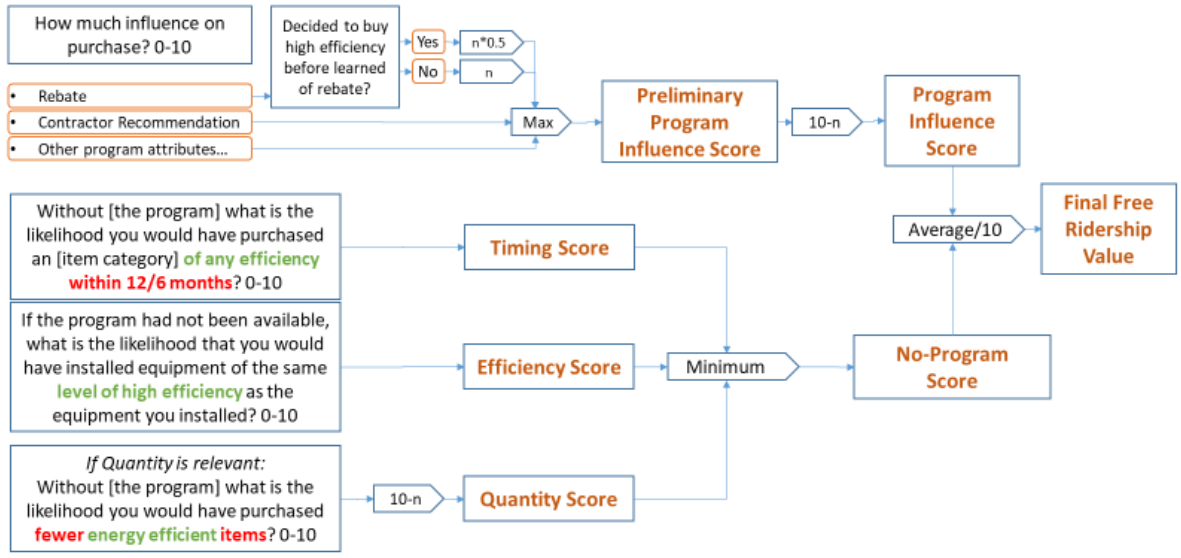
The battery includes two main free-ridership components. Each of these components assesses the likelihood of free-ridership on a scale of 0 to 10, and the two scores are averaged for a combined total free-ridership score.

1. A Program Influence component – this is based on the participant's perception of the program's influence on carrying out the energy-efficient project
2. A No-Program component – this is based on the participant's intention to carry out the energy-efficient project without program funds.

Figure 2 from the IL TRM illustrates the scoring algorithm.

[51] Respondents are asked to report their likelihood (using a 0 to 10 scale, where 0 is *not at all likely* and 10 is *extremely likely*) to implement specified energy-efficiency measures in the absence of the program. That likelihood score is then divided by 10 to produce the no-program score.

Figure 2: IL TRM Residential Prescriptive Rebate (with no Audit) Free-Ridership Algorithm



In the figure above, the number of program elements (e.g., rebate, contractor recommendation, other program attributes) will vary depending upon the program's design.

The No-Program Score is the minimum of the Timing, Efficiency, and (if applicable) Quantity Scores. The No-Program Score is averaged with the Program Influence Score to calculate the Final Free-Ridership Score, as shown in the following algorithm:

No Program Score (NP) = Min (T, E, Q)

Free-Ridership (FR) = Mean (PI, NP)

The protocol further specifies that the survey should include open-end consistency checks to address the possibility of conflicting responses in the two scores. If the inconsistency exists and the open-ended response does not resolve the inconsistency, the respondent should be removed from the calculation. If questions have missing responses, the case should be retained in the analysis.

A.3.1.2 The Single-Family Home Energy Audit

This protocol is for programs that provide audits or energy assessments. These programs may also include direct-install measures and incentives for energy-efficiency opportunities through the program or other PA-sponsored programs. Free-ridership and participant spillover estimates rely on participant self-reports.

The basic method specifies that because audit programs may have multiple components, net impacts should be estimated using survey batteries tailored to a customer's experience (e.g., receipt of free direct-install measures and discounted or rebated measures).

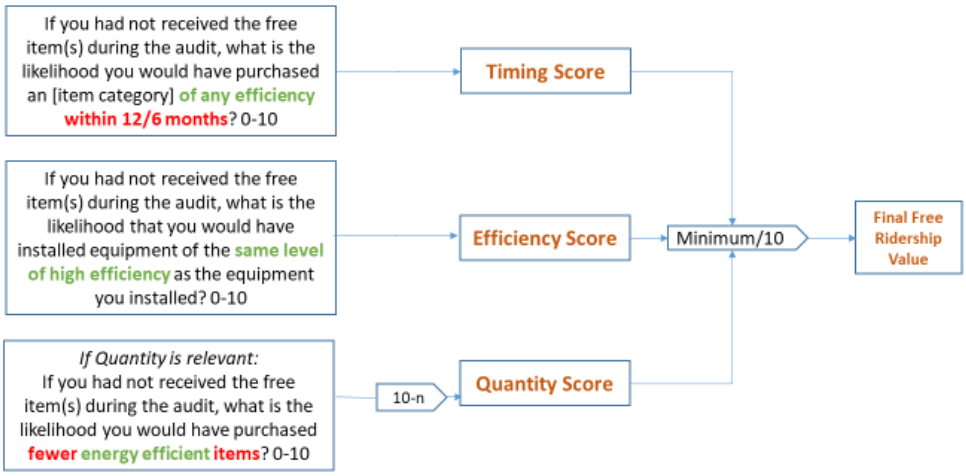
For no-cost direct install measures, the free-ridership calculations include questions on Timing, Efficiency, and Quantity. Due to the low cost of the measures and the number of measures, the

protocol provides a streamlined battery of questions to reduce respondent burden. The algorithm is the minimum of the timing, efficiency, and quantity scores.

Free-Ridership (FR) = Min (T, E, Q)

Figure 3 from the IL TRM illustrates the battery.

Figure 3: IL TRM Single-Family Home Energy Audit Free-Ridership – No Cost Measures



For low or no-cost direct-install measures, two consistency questions should be used. The first is asked at the beginning of the FR battery, and the second is at its conclusion. If the inconsistency exists and the open-ended response does not resolve the inconsistency, the respondent should be removed from the calculation. If questions have missing responses, the case should be retained in the analysis. The questions are as follows:

- Prior to the audit, had you purchased any <measures>? Y/N
- IF YES AND LIKELIHOOD TO INSTALL WITHOUT THE PROGRAM IS <7: Given that you had purchased <measures> before receiving the audit, why didn't you purchase additional <measures> on your own without the program?

IF NO AND LIKELIHOOD TO INSTALL WITHOUT THE PROGRAM IS >6: Given that you had NOT purchased <measures> before receiving the audit, why were you likely to purchase <measures> on your own without the program?

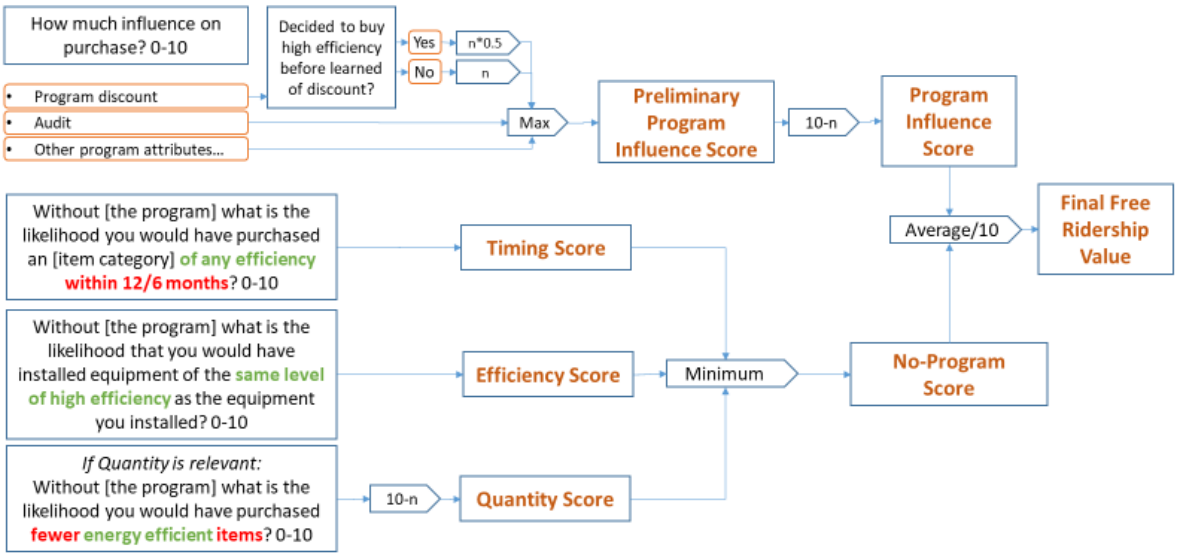
For rebated/discounted measures (typically for building shells), the protocol measures FR using a customer SRA with questions on two components:

- A Program Influence component – this assesses the participant's perception of program elements including the discount and the audit itself on carrying out the energy-efficiency project

A No-Program component – this is based on the participant's likelihood of purchasing he exact same items at the same time in the absence of the program.

This method is identical to that used in the Prescriptive Rebate (with no audit) protocol with the one exception that the questions about program influence must include the audit itself as one of the program attributes. The algorithm from the IL TRM is presented in Figure 4.

Figure 4: IL TRM Single-Family Home Energy Audit Free-Ridership – Discounted Measures



The protocol further specifies that the survey should include open-end consistency checks to address the possibility of conflicting responses in the two scores. If the inconsistency exists and the open-ended response does not resolve the inconsistency, the respondent should be removed from the calculation. If questions have missing responses, the case should be retained in the analysis.

A.3.1.3 The Multifamily Protocol

The Multifamily protocol is used for programs that offer direct installation of low-cost, energy-efficient measures in multifamily dwelling units, in addition to rebates for common areas. The basic method for estimation of free-ridership and PSO for these types of programs is based on participant SRA. For common area and building shell components of the program, participants are property managers and owners responsible for building maintenance and renovation. Evaluators may field surveys with owners, property managers, or tenants, depending on the program's design and theory.

The free-ridership method for discounted measures is identical to that used in the Prescriptive Rebate (with no Audit) protocol with one exception. The questions about program influence should be sure to include the audit itself as one of the program attributes. As shown in Figure 5 and Figure 6, the IL TRM contains an algorithm for non-CFL/non-LED measures and one for CFL/LED measures.

Figure 5: IL TRM Multifamily Free-ridership – Non-CFL/Non-LED Measures

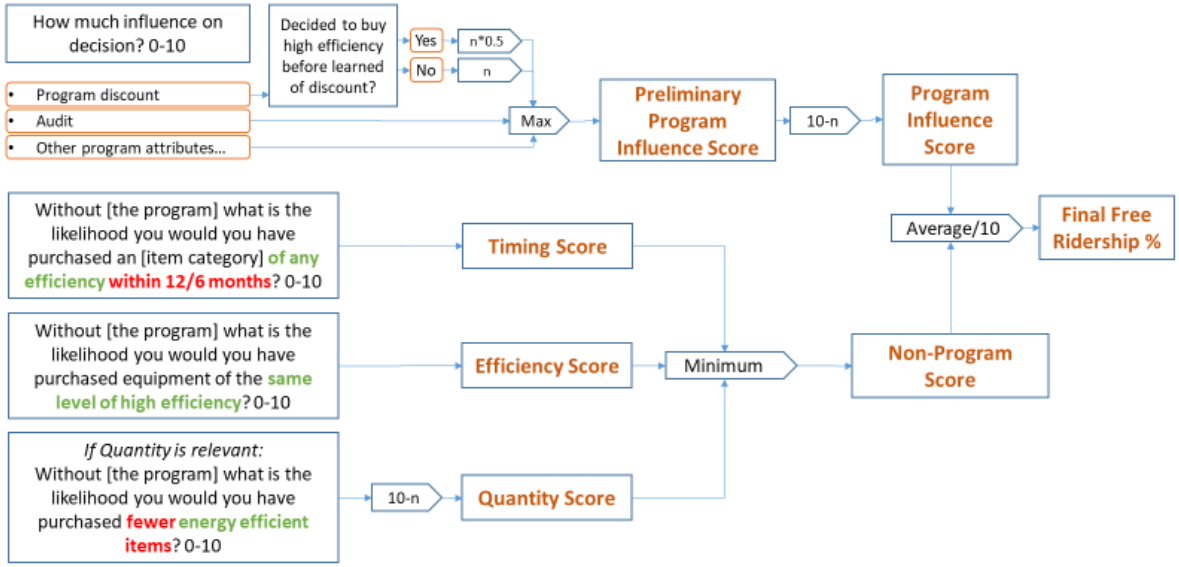
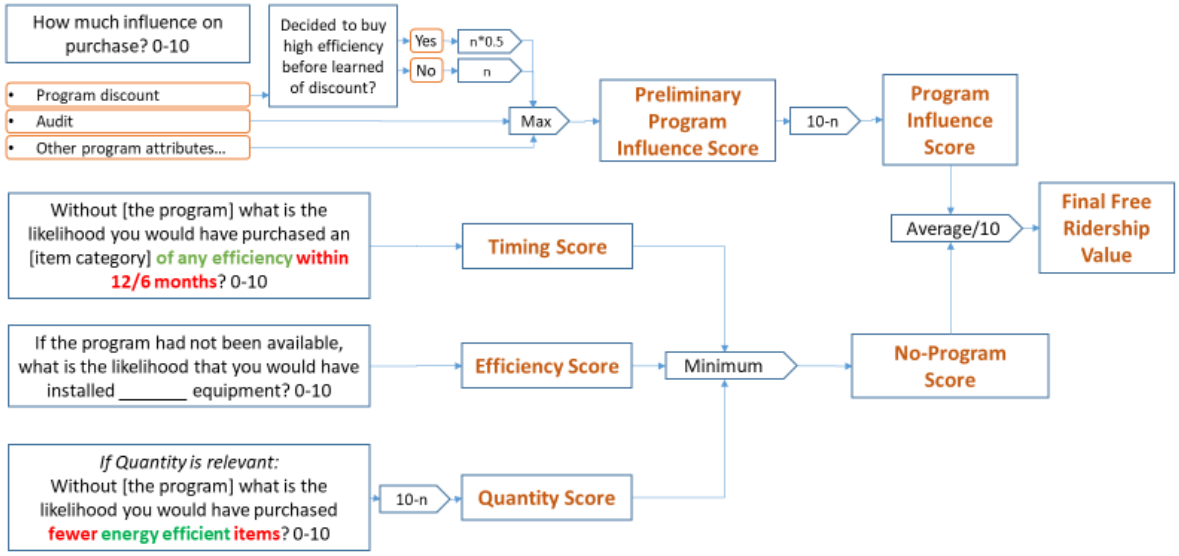


Figure 6: IL TRM Multifamily Free-ridership for Property Owners – CFL/LED Measures



The protocol further specifies that the survey should include open-end consistency checks to address the possibility of conflicting responses in the two scores. If the inconsistency exists and the open-ended response does not resolve the inconsistency, the respondent should be removed from the calculation. If questions have missing responses, the case should be retained in the analysis.

A.3.1.4 The Energy Saving Kits and Elementary Education Protocol

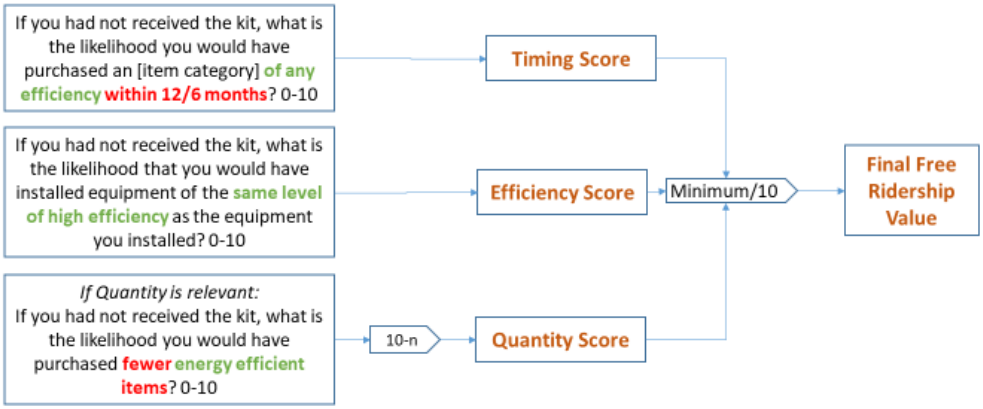
This protocol is used for programs that distribute kits containing various energy-saving measures. Customers can request a kit, or it may be delivered through an elementary education program. Free-ridership and PSO estimation for both delivery methods rely upon participant self-report information from surveys.

The basic free-ridership calculations include Timing, Efficiency, and Quantity components. The final free-ridership score is calculated by taking the minimum of the Timing, Efficiency, and quantity scores, as follows:

Free-Ridership (FR) = Min (T, E, Q)

Figure 7 from the IL TRM describes the method.

Figure 7: IL TRM Energy Saving Kits and Elementary Education Free-Ridership



Free-ridership should be calculated for each separate kit component, and then weighted by savings to determine program-level results. Missing responses to specific questions should be treated as missing for that particular question, but the case should be retained in the analysis.

A.3.1.5 The Residential New Construction Protocol

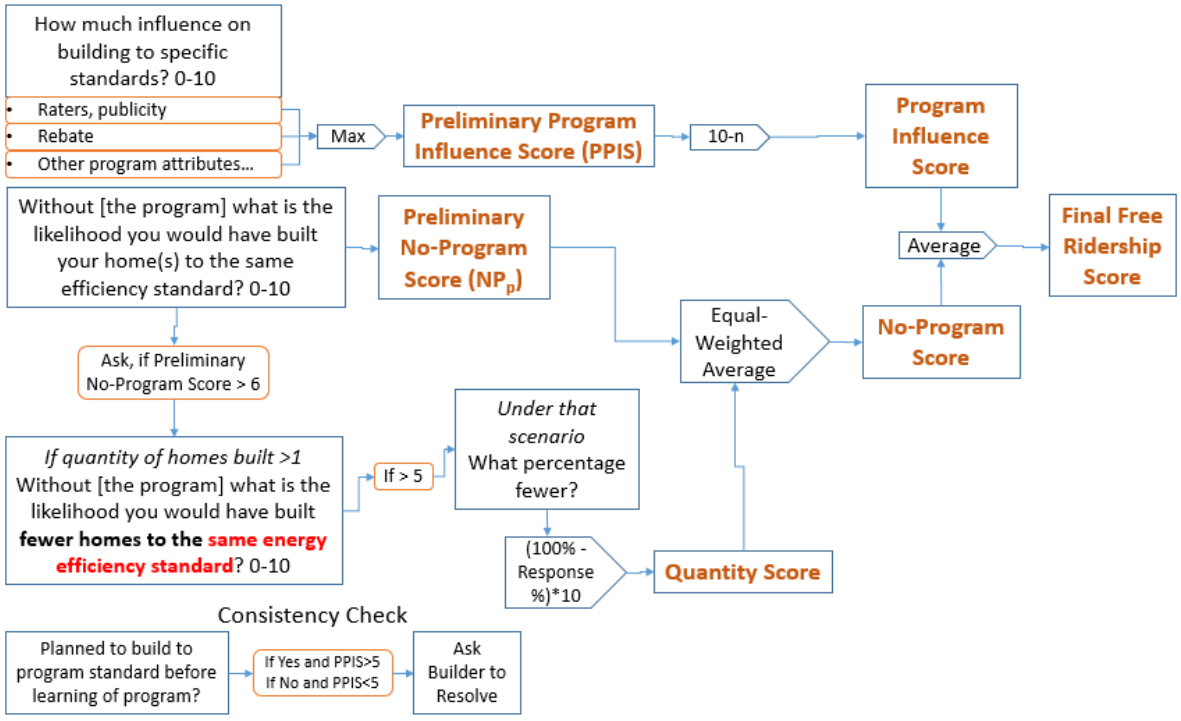
This protocol uses a basic method for estimating free-ridership and participant spillover that is based on builder participant SR through surveys. It is used for programs that typically offer builder training, technical information, marketing materials, and incentives to builders for the construction of eligible homes.

For this type of program, a free rider is a builder who would have constructed a home at the program's efficiency level in the program's absence. Survey questions consider the builder's likelihood of meeting the same energy-efficiency standard, rather than whether or not the builder would have installed certain energy-efficiency measures. Evaluators assess Program Influence by asking respondents, on a scale from 0 (not at all important) to 10 (extremely important), how important they found various program elements in deciding to build to specific energy-efficiency standards. The number of elements included vary, depending on the program's design. Logic

models, program theory, and staff interviews typically inform the list of program elements included. Programs typically use the following elements to influence builder actions: marketing materials; incentives or rebates; contacts with HERS Raters; and technical assistance.

In addition to asking about specific program influences, surveys should ask builders whether they planned to build homes to the same standard before learning of the program. Figure 8 presents the new construction algorithm.

Figure 8: IL TRM Residential New Construction Free-Ridership Protocol



The Program Influence Score (PI) equals 10 minus the maximum influence rating for any program element rather than, for example, the mean influence rating. This is based on the rationale that if any given program element had a great influence on the respondent's action, the program itself had a great influence, even if other elements had less influence.

Evaluators should calculate the No-Program score using a set of questions that ask respondents to gauge their likelihood of building homes to the same standards and in the same quantities had the program not existed. Three separate responses are considered in calculating the No-Program Score:

- The likelihood, on a scale of 0 to 10, that the builder would have built their homes to the same efficiency standard (Preliminary No-Program Score [NPp])
- If that likelihood is greater than 6, the likelihood of fewer homes being built to the same efficiency standard.

- If that likelihood is greater than 6, the response to the question "for that scenario, what percentage of fewer homes would be built to the standard?" (Quantity Score = [100% - % answer] * 10, which will be a number between 0 and 10)

The resulting No-Program (NP) Score is calculated as follows:

$$NP=Mean(NP_p,Q)$$

The overall Free-Ridership Score is the average of the PI and NP scores:

$$FR=Mean\ (PI,\ NP)$$

A.3.1.6 The Appliance Recycling Protocol

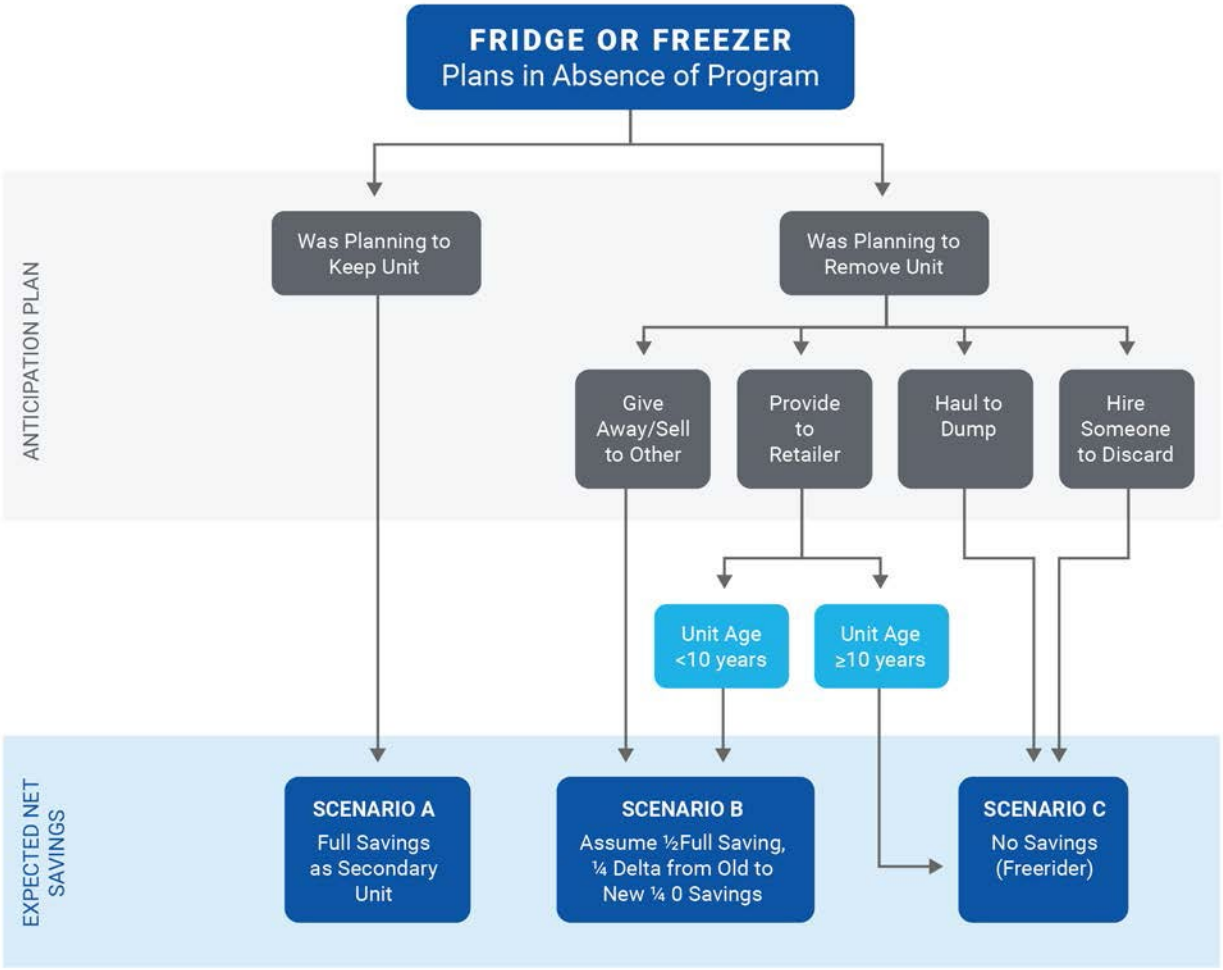
The free-ridership estimation approach for the Appliance Recycling Protocol significantly differs from other approaches since the program theory and logic for appliance recycling differs significantly from standard downstream incentive programs. This protocol was adapted from the Pennsylvania Statewide Evaluator Common Approach for Measuring Net Savings for Appliance Retirement Programs.[52]

The protocol includes basic and enhanced SRA methods with a multistep process to segment participants into different groups, each with specific attributable savings in the absence of program intervention: (1) the appliance would have been kept by the participating household, (2) the appliance would have been discarded in a way that transfers the unit to another customer for continued use, and (3) the appliance would have been discarded in a way that would have permanently removed the unit from service. Only the third option constitutes free-ridership. Options 1 and 2 indicate non-free riders, although additional questions are needed to further classify these participants for secondary market impacts.

For the basic method, the protocol recommends a participant survey for estimating NTG. Figure 9 from the IL TRM illustrates how participants fall into one of the three groups based on provided survey questions.

[52] Guidance Memo-026, March 14, 2014.

Figure 9: IL TRM Appliance Recycling Protocol



The scoring algorithm is as follows:

FR=(free-ridership and secondary market impacts % - induced replacement %)

The results of the basic approach can be enhanced with three additional research efforts. Note that the basic method has defaults for where primary research on enhanced approaches cannot be performed. As described in the IL TRM, the three additional research efforts include the following:

1. A retailer survey to determine the proportion of units returned to a retailer and that the retailers would recycle.
2. An appliance market assessment study to determine the size of the secondary appliance market and whether removal of participating units from the market would cause an otherwise would-be receiver to purchase an alternative used or new unit.
3. A non-participant survey to assess how non-participants acquire and dispose of used units.

The protocol does not provide specific guidance for when to use each SRA method nor for measuring PSO and NPSO, and thus the cross-cutting protocols may be assumed to prevail.

A.3.2 Participant and Non-Participant Spillover with Customers and Trade Allies

The Illinois protocols include two general methods of assessing spillover, one through end-user (or participant/non-participant) research and the other through trade ally research. If end-user research is combined with trade ally research, there is a potential for overlap in the resulting spillover estimates, so care must be taken to avoid double-counting.

A.3.2.1 Participant Spillover Measured from Customers

The survey should first ask about the influence the program had on their taking additional energy-saving actions on their own, then follow-up with two close-ended questions. The questions are as follows:

1. Did the program influence you in any way to make these additional improvements?
2. How important was your participation in the <Program Administrator's> program on your making additional energy-efficiency improvements on your own? (0 to 10 scale, where 0 means *not at all important* and 10 means *extremely important*.)
3. If you had not participated in the <Program Administrator'> program, how likely is it that you would still have implemented this measure, using a 0 to 10 scale, where 0 means you definitely WOULD NOT have implemented this measure and 10 means you definitely WOULD have implemented this measure?

The response to Q. 2 is Measure Attribution Score 1, and the response to Q. 3 is Measure Attribution Score 2. The measures referenced in the question are considered to be attributable to the program if the Spillover Score is greater than 5.0:

Spillover Score = (Measure Attribution Score 1 + (10-Measure Attribution Score2))/2>5.0

Therefore, energy impacts associated with measures implemented outside the program are either 100% program-attributable or 0% program-attributable.

If the above conditions are met, customers are asked how they know the measure is more efficient than other models. If the respondent identifies the measure as ENERGY STAR or names an efficiency level that the evaluator confirms is above the minimum federal standards, it counts towards Participant spillover.

Finally, depending on the measure type, follow-up questions should ask customers to provide reasonable information to allow the evaluator to estimate the amount of savings using the IL TRM protocols (e.g., quantity or location and amount of insulation. The algorithm is as follows:

Participant Spillover Rate (PSO) = Sum of energy or demand from additional measures installed/sample ex post gross energy or demand impacts

The TRM recommends the sample be of current or up to one year of previous program participants.

A.3.2.2 Non-participant Spillover Measured from Customers

If the evaluation measures non-participant spillover from customers, this will require a large sample of customers who have not participated in any energy-efficiency or behavioral programs within the past three years. Also, care must be taken to ensure spillover is not double-counted with a trade-ally approach.

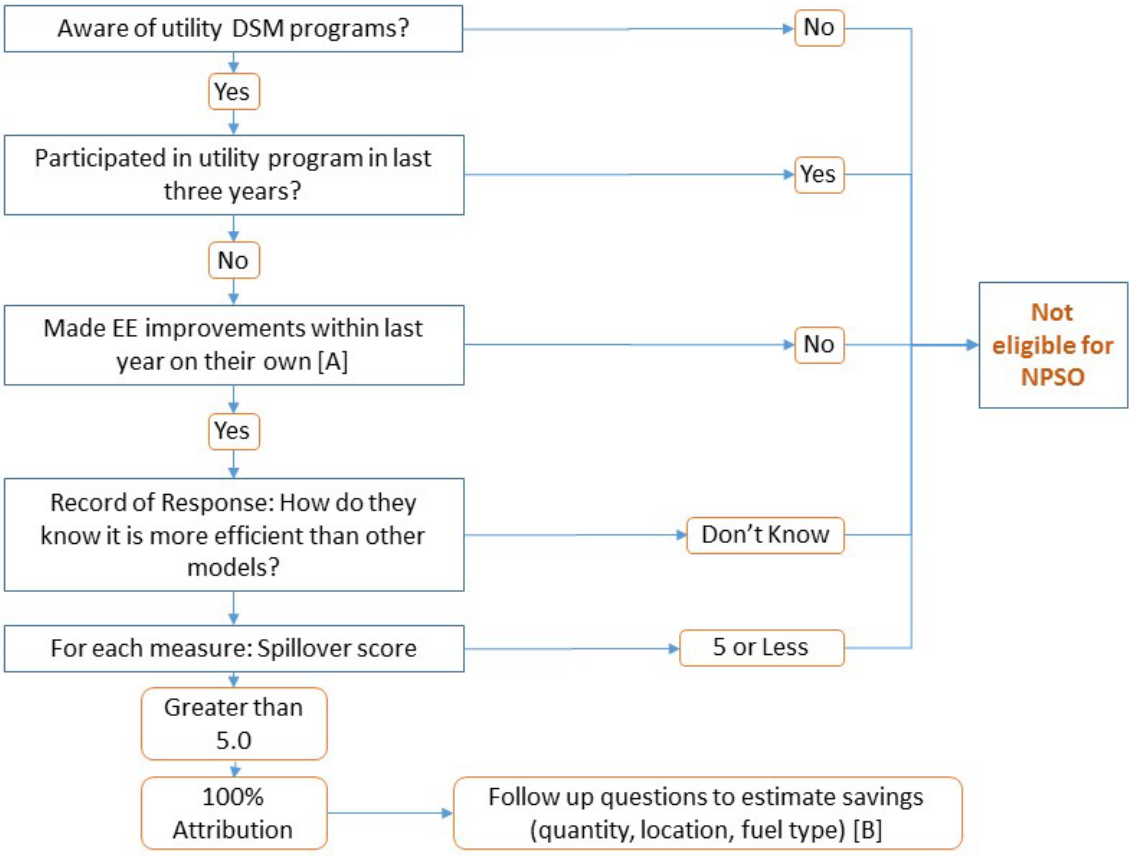
The basic method uses a two-step process: (1) conduct a non-participant survey to identify potential spillover measures, and (2) if needed, conduct a follow-up call or on-site visit by technical staff to confirm attribution and obtain information needed to estimate energy savings.

Depending on the spillover measure type reported by the customer, follow-up questions should be included to gather enough information to reasonably assess the saving amount by applying the IL TRM, understanding that assumptions must be made if IL TRM inputs cannot be easily supplied by the participant. Baselines for measures not in the IL TRM will be assessed based on appliance standards and building codes, if applicable, and, if not, through engineering judgements of existing or market conditions. Engineering assumptions and analysis by the evaluator will be applied for measures not included in the IL TRM. Key assumptions should be documented in the report.

To receive credit for energy savings, the non-participant must fit the following criteria: (1) be familiar with the PAs energy-efficiency campaign and (2) indicate that some aspect of the PA's energy-efficiency programs motivated their purchases. Influence will be measured on a scale of 0 to 10, where 10 is extremely influential and 0 is not at all influential. As shown in Figure 10, savings attribution requires a Spillover Score of greater than 5.0.

Spillover Score = (Measure Attribution Score 1 + (10-Measure Attribution Score2))/2>5.0

Figure 10: IL TRM NPSO Question Logic



A.3.2.3 Non-participant Spillover Measured from Trade Allies

The protocol provides approaches for measuring spillover from both active and nonactive trade allies.

For downstream programs, the protocol recommends surveys of active trade allies to see if the program influenced their sales of high-efficiency equipment and to quantify the program's impact on their high-efficiency sales. To assess if a sampled trade ally created spillover, the following screening criteria are recommended (the order of these may be adjusted by the evaluator):

1. Percentage of trade ally's installations/sales that are high efficiency and/or the total volume of high-efficiency installations/sales increased since the trade ally became exposed to the program.
2. The trade ally rated the program as important to at least one of these high-efficiency installation increases.
3. The trade ally installed/sold at least some high-efficiency equipment or products that did not receive an incentive.
4. The trade ally's recommendation was influential in the customers' choice of high-efficiency equipment.

5. The open-ended response about why customers with eligible projects do not receive an incentive supported that the non-incented high-efficiency installations can be considered spillover.

Must pass one of the above screening criteria to qualify for spillover. To quantify spillover, the survey collects information on the percentage of the trade ally's total equipment installations/sales that was (1) standard efficiency, (2) high efficiency that did receive a program incentive, and (3) high efficiency that did not receive a program incentive.

% of TA's High-efficiency Equipment that Received Incentive = *% High efficiency that DID receive a program incentive* ÷ *% High efficiency that DID receive a program incentive* + *% High efficiency that did NOT receive a program incentive*

Savings of Non-Incented High Efficiency Equipment = *Savings from Program Database* ÷ *% of TA's high-efficiency equipment that received incentive* - *Savings from Program Database* * *Size Adjustment*

A.3.3 Participant and Non-Participant Spillover with Builders (Residential New Construction)

Participant spillover is calculated based on participant builder survey questions that ask builders about homes built within the utility service territory but outside the program. Survey questions ask whether the builder increased the energy-efficiency standards of non-program homes after participating in the program, and the number of homes they applied these increased standards to, within the utility's service territory. Depending on the program characteristics, spillover should be measured as changes in specific building practices or as installation of specific measures. The text below from the TRM assumes the program targets modifying building practices.

Spillover may be recorded depending on responses to the following questions:

1. How important was your experience in the <PROGRAM ADMINISTRATOR'S> program in your incorporating this building practice your other homes, using a scale of 0 to 10, where 0 is not at all important and 10 is extremely important?
2. If you had not participated in the <PROGRAM ADMINISTRATOR'S> program, how likely is it that you would still have incorporated this building practice using a 0 to 10, scale where 0 means you definitely WOULD NOT have implemented this practice and 10 means you definitely WOULD have implemented this practice?

Responses to the first question establish the Practice Attribution Score 1, and responses to the second question establish the Practice Attribution Score 2. Spillover may be program-attributable for building practices with self-report data meeting the following condition:

Spillover Score = (Practice Attribution Score 1 + (10 – Practice Attribution Score 2))/2 > 5.0

For responses meeting these conditions, an evaluator determines that specific building practices referenced in the question are attributable to the program; otherwise, the evaluator determines that specific building practices referenced in the question are not attributable to the program. The attribution criteria represent a threshold approach, in which energy impacts associated with building practices program participants implement outside the program are either 100% program-attributable or 0% program-attributable.

For each building practice discussed, builders will be asked how they know the building practice is more efficient than other options. If the respondent can identify the building practice as ENERGY STAR or name an efficiency level that the evaluator confirms as above the minimum federal standard, or if they identify a technology that the evaluator can confirm is above the minimum federal standard, this counts towards participant spillover.

Finally, depending on the building practice cited by the builder, follow-up questions should ask customers to provide reasonable information to allow the evaluator to estimate the amount of savings using IL TRM protocols, such as quantity of appliances or the location and amount of insulation.

To calculate the spillover energy and demand savings for these actions, further questions should be asked to assess the gross savings of the building practice, through the appropriate version of the IL TRM, if available, and the number of homes to which it applied. To develop the Spillover Rate, the total energy and demand impacts from the sampled participants who implemented efficient building practices in other homes due to participation in the program is summed, and then this sum is divided by the total ex post sample energy and demand impacts.

The equation used to calculate the Core NTGR based on participant spillover is as follows:

NTG=(1-FR+PSO)

Builder non-participant spillover occurs when non-participating builders implement some or all of the efficiency measures incorporated through the program in order to compete with builders that are participating. Non-participant spillover is determined by surveying (1) drop out builders who previously participated in the program over 12 months ago and (2) true non-participating builders who are aware of the program or aware of other builders who are taking steps to improve new home efficiency.

The survey questions first identify specific building practices that go beyond the energy code in the builder's jurisdiction. For each practice that is more efficient than code, the survey asks additional questions on how many homes the guilder sold that incorporated this upgrade, and of these homes, how many would have incorporated this upgrade.

A.4 New York Clean Energy Guidance[53]

This Clean Energy Guidance document provides guidance to the utilities and the New York State Energy Research and Development Authority and evaluators on how to conduct evaluation, measurement and verification (EM&V) activities associated with ratepayer funded clean energy programs. The guidance document includes guidance on NTG for Conventional Resource Acquisition Program Impact Evaluation (see Appendix F of the Guidance). The guidance in this appendix covers the following:

1. Program-Level Participant and Non-Participant Spillover
2. Estimating NTG Ratios Using SRA
3. Calculating the Relative Precision of Program Net Savings

Free-ridership and spillover are captured in the NTG ratio to reflect the degree of program-induced actions. The equation for the NTG ratio is as follows:

NTG ratio = (1-Free-ridership) + Spillover (PSO and NPSO)

The remaining sections are largely quoted from New York Clean Energy guidance document.

A.4.1 Spillover

The guidance suggests that there might be situations in which all key stakeholders are willing to agree that spillover is not zero, but the expense to estimate it reliably is prohibitive. In these cases, PAs may utilize a deemed spillover rate that is based on a review of the literature and the program theory and logic model, that together describe reasonably well the causal mechanism that is expected to generate spillover.

[53] http://www3.dps.ny.gov/W/PSCWeb.nsf/96f0fec0b45a3c6485257688006a701a/255ea3546df802b585257e38005460f9/$FILE/CE-05-EMV%20Guidance%20Final%20%2011-1-2016.pdf

A.4.1.1 Program-Level Participant and Non-participant Spillover

The guidance defines spillover as follows:

. . . the energy savings associated with energy-efficient equipment installed by consumers who were influenced by an energy-efficiency program, but without direct financial or technical assistance from the program. Spillover includes additional actions taken by a program participant, as well as actions undertaken by non-participants who have been influenced by the program.

This definition is consistent with the somewhat more detailed definition contained in the California Energy Efficiency Policy Manual (2008):

> Reductions in energy consumption and/or demand in a utility's service area caused by the presence of the DSM program, beyond program related gross or net savings of participants. These effects could result from (1) additional energy-efficiency actions that program participants take outside the program as a result of having participated; (2) changes in the array of energy-using equipment that manufacturers, dealers, and contractors offer all customers as a result of program availability; and (3) changes in the energy use of non-participants as a result of utility programs, whether direct (e.g., utility program advertising) or indirect (e.g., stocking practices such as (2) above or changes in consumer buying habits). Participant spillover is described by (1), and non-participant spillover by (2) and (3). Some parties refer to non-participant spillover as *free-drivers*. (TecMarket Works Team, 2006)

Some evaluators subdivide participant spillover into *inside* and *outside* spillover. Inside spillover occurs when, due to the project, additional actions are taken to reduce energy use at the same site, but these actions are not included as program savings. Outside spillover occurs when an actor participating in the program initiates additional actions that reduce energy use at other sites that are not participating in the program. Because causality is inherent in the very definition of spillover, the spillover savings are inherently net.

Free-ridership and spillover are captured in the NTG ratio to reflect the degree of program-induced actions. Specifically, the gross energy savings estimate, refined by the realization rate, is adjusted to reflect the negative impacts of free-ridership and the positive impacts of spillover. Equation 1 illustrates this adjustment.

$$NTG\ ratio = (1\text{-}Free\text{-}ridership) + Spillover$$

Clearly, ignoring spillover results in a downward bias in the NTG ratio.

A.4.1.2 Key Decisions for Evaluators

Before evaluators decide to estimate spillover, they must make a number of critical decisions:

1. Will the evaluation address participant spillover, non-participant spillover, or both?
2. Does the size of the expected savings warrant the expenditure of evaluation funds needed to estimate these savings at an appropriate level of reliability?
3. Which of the two levels of methodological rigor discussed in these guidelines, standard or enhanced, should be used?
4. Will spillover be estimated based on data collected from end users, those upstream from end users (e.g., vendors, installers, manufacturers), or both?
5. What is the level of aggregation? Although participant spillover is always estimated at the program level, if an evaluator is attempting to estimate non-participant spillover, will the evaluator estimate it at the program level or the market level? One potential reason for estimating non-participant spillover at the market level is that, in some circumstances, reliably teasing out the spillover savings attributable to one specific program among many may be nearly impossible due to the difficulty non-participants may have in attributing any of their installations to a specific program. In such a case, evaluators can choose to conduct market effects studies, which include naturally occurring adoptions, program-rebated adoptions, participant and non-participant spillover, other program effects that cannot be reliably attributed to a specific program (e.g., upstream lighting programs and the effects of the portfolio of programs on such things as increases in the allocation of shelving space to efficient measures), and other non-program effects due to such factors as DOE Energy Star; programs funded by the American Recovery and Reinvestment Act (ARRA); and the gradual non-program induced evolution of the market in terms of attitudes, knowledge, and behavior regarding energy efficiency. The net savings resulting from market effects studies must be included in the portfolio-level benefits-costs analyses.
6. If an evaluator decides to conduct a market effects study, then they must decide whether the study should be focused on the region targeted by a given PA, multiple regions, or even the entire state.

Once these questions are answered, evaluators can then use these guidelines in estimating spillover.

A.4.2 Estimating Net-To-Gross Ratios Using the Self-Report Approach

The SRA guidelines require analysts to address certain key issues but does not require analysts to address these issues in a specific way. According to the guidelines, “The SRA is useful in a variety of situations. For example, in some cases, the expected magnitude of the savings for a given program might not warrant the investment in an expensive evaluation design that could involve a billing analysis or a discrete choice analysis of both participants and non-participants and that would address self-selection bias. Or, key stakeholders might not want to wait for a billing analysis to be completed. Also, if the relationship of the savings to the normal monthly variation in energy use is too small, then a billing analysis should not be attempted owing to a lack of statistical power. Finally, in some cases, it might not be possible to identify a group of customers to serve as a comparison group, since they have been exposed through prior participation or are

in some other ways *contaminated*. So, for budgetary, timing, statistical, and research design issues, the more traditional designs and analyses must sometimes be replaced with the SRA."

The SRA is a mixed method approach that involves asking one or more key participant decision-makers a series of structured and open-ended questions about whether they would have installed the same energy-efficient equipment in the absence of the program. It also includes questions that attempt to rule out rival explanations for the installation. The New York guidance does not specify questions or algorithms but provides guidance on best practices for evaluators to use.

Timing of the SRA data collection

In order to minimize the problem of recall, SRA interviews addressing free-ridership should be conducted with the decision maker(s) as soon after the installation of equipment as possible, while interviews or other data collection to assess spillover need to be conducted later to allow enough time for the occurrence of spillover.

Set-up Questions

An evaluation using self-report methods should employ and document a set of questions that adequately prompt the respondent to the context and sequence of events that led to decision(s) to adopt a DSM measure or practice, including clearly identified benchmarks in the customer's decision-making process.

Decision-Making Process

A key purpose of the decision-making questions is to help the respondent recall the particulars of their program-related decision-making and prepare them to answer the direct attribution questions about how the program affected the timing, efficiency level, and quantity of the technology installed.

Ruling out Rival Hypothesis

Evaluator should attempt to rule out rival hypotheses regarding the reasons for installing the efficient equipment.

Consistency Checks

An evaluation using self-report methods should employ a process for setting up checks for inconsistencies when developing the questionnaire items, and describe the methods chosen, as well as the rationales for using or not using the techniques for mitigating inconsistencies.

Early Replacement Considerations

Early replacement is defined as the replacement of equipment before it reaches its Effective Useful Life (EUL), whereas end-of-life or normal replacement refers to the replacement of equipment that has reached or passed the end of its measure-prescribed EUL. For projects associated with early replacements, the evaluator must first verify that a given installation is actually a case of early replacement. The guidance suggests asking participants the following types of questions to identify early replacement. If it is deemed that it is early replacement, then the guidance provides eight variables and associated documentation to make necessary adjustments to gross savings:

- Approximately how old was the existing equipment?
- How much longer do you think it would have lasted?
- In your opinion, based on the economics of operating this equipment, for how many more years could you have kept this equipment functioning?
- Which of the following statements best describes the performance and operating condition of the equipment you replaced through the PROGRAM?
    - Existing equipment was fully functional
    - Existing equipment was fully functioning, but with significant problems
    - Existing equipment had failed or did not function
    - Existing equipment was obsolete
    - N/A/ (e.g., VSD, EMS, controls)
- How much downtime did you experience in the past year, and how did this compare with the previous year(s)?
- Over the last five years, have maintenance costs been increasing, decreasing or staying about the same.

A.4.3 Other NTG approaches

The Guidance presents a number of other NTG approaches that could be used and discusses that the selection of methods for use in evaluation of a given program will depend on a number of factors. These other methods include the following:

- Theory Based Evaluation
- Experimental designs
- Quasi-experimental designs
- Price elasticity approaches, including conjoint analysis and revealed preference analysis
- Structured expert judging
- Historical Tracing: Case Study Method
- Top-Down Modeling

Primary suggested approaches for residential programs are listed below (secondary in parenthesis).

- Whole building retrofits = market actor self-reports, theory based evaluation, quasi-experimental
- HER = theory based evaluation, experimental design, quasi-experimental (market actor self-reports)
- Upstream = theory based evaluation, quasi-experimental, price elasticity/econometric (market actor self-reports, expert judging)
- New Construction = market actor self-reports, theory based evaluation, expert judging (quasi-experimental, case studies).

A.5 California Guidance Documents

The California Evaluation Framework, developed in 2004 for the California Public Utilities Commission and the Project Advisory Group,[54] provides guidance on different methods that can be used to assess NTG, including the SRA, but it does not prescribe survey questions or algorithms. The Framework notes that the SRA involves asking one or more key participant decision-makers a series of closed and open-ended questions about their motivations for installing the efficient equipment and the efficiency of the equipment that they would have installed in the absence of the program. Additional questions are typically asked to establish the temporal precedence of the program and to rule out rival explanations for the installation.

The Framework states that adding a consistency check question and adjusting individuals estimates accordingly also improves the estimate. The Framework recommends dropping respondents with inconsistent responses rather than assuming central estimates for them.

The Framework recommends developing a standardized survey method for use in California, noting that a carefully constructed survey with checks and triangulation methods might offer a more *accurate* measurement of a program's free-ridership. They suggest two improvements to the Massachusetts C&I standardized batteries created in 2003: (1) developing a method for using contingency questions and for dropping responses from inconsistent respondents from the analysis if the inconsistency cannot be corrected, and (2) testing and analysis to calibrate the survey against a mix of econometric methods.

In terms of timing, the Framework notes that the best available estimate of the NTGR is the latest estimate for a program. The Framework also recommends that deemed NTGR should never be used to report net savings within an evaluation.

In addition to the California Framework, guidelines for estimating NTG ratios using the SRAs were established.[55] The SRA deviates from the standard approach to assessing causality (experiment or quasi-experiment), which is not always desirable or possible. The SRA guidelines are as follows:

1. Timing of the interview: this should be as soon after installation as possible.
2. Identifying the correct respondent(s: this is critical, especially for large C&I situations where different actors have different and complementary pieces of information about the decision making; decisions may be being made in regional or national headquarters; decision making may be done by commissions, committees, boards or councils; and there may be both a technical and a financial decision maker.
3. Set up questions: these are needed to adequately establish the context and sequence of events that led to decisions.

---

[54] The California Evaluation Framework, (2004), Prepared for the California Public Utilities Commission and the Project Advisory Group.[54] https://www.cpuc.ca.gov/uploadedFiles/CPUC_Public_Website/Content/Utilities_and_Industries/Energy/Energy_Programs/Demand_Side_Management/EE_and_Energy_Savings_Assist/CAEvaluationFramework.pdf

[55] Ridge, Richard, Ken Keating, Lori Megdal, and Nick Hall. *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches.* October 15, 2007.

4. Use of multiple questions: using both quantitative and qualitative questions to measure a construct, such as free-ridership, is preferable as it increases reliability.
5. Validity and reliability: These should be assessed for each question used. The internal consistency of multiple-item scales should also be tested. For large savings sites, multiple members of the evaluation team should review the results to ensure consistency.
6. Consistency checks: set up checks for inconsistencies and establish rules for handling inconsistent responses while the respondent is on the phone.
7. Measure-specific questions: the questions should be measure-specific.
8. Partial free-ridership: explore cases where participant would have installed something more efficient than program-assumed baseline but not as efficient as program equipment.
9. Deferred free-ridership: Measure the program's impact on acceleration of installation. Should use more than one question and use preponderance of evidence approach.
10. Scoring algorithms: This can impact results and must be documented and tested using sensitivity analyses. A preponderance of evidence approach is better than relying solely on an algorithm.
11. Handling "don't knows" and non-responses: determine in advance how these will be handled. Make a special effort to avoid "don't know" responses.
12. Weighting the NTGR: take into account the size of the savings impacts at the customer or project level.
13. Ruling out rival hypotheses: ask open-ended questions regarding other possible reasons for installing the efficient equipment.
14. Precision of the estimated NTGR: This should be reported but it is complicated when there are multiple sources of information or multiple respondents. In such cases, to take into account the propagation of errors in the relative precision.
15. Pretesting the instrument: always pretest to reveal ambiguous wording, faulty skip patterns, leading questions, faulty consistency checks, and incorrect sequencing of questions.
16. Large Savers, complex decision making: In cases of large savers or complex decision-making, incorporate additional quantitative and qualitative data when estimating the NTGR. For example,
    a. use multiple respondents
    b. use other site- and market-level data
    c. establish rules for data integration
    d. consider the analysis method. Case studies are one method for assessing both quantitative and qualitative data, content analysis to identify coherent and important these and patterns in the data. Use of multiple evaluators to independently review the data.
17. Qualified interviewers: for complex situations, engineers familiar with the more complicated technologies should be trained to collect the data.

The guidelines note the importance of where the free-ridership and spillover questions are placed in the survey instrument. Before the questions, evaluators need to have *warm-up* questions to enable better respondent recall of their decision-process for the measures/products being discussed. The placement must also ensure that none of the questions prior to the self-report batteries for free-ridership and participant spillover would be likely to bias the responses. The guidelines suggest including these questions immediately after the installation verification questions. When a measure has been removed, the free-ridership questions will not be asked of that measure.

The Energy Division convened a committee of experienced evaluators to develop a standard framework for the measurement and calculation of NTG ratios for residential and small commercial programs in a systematic and consistent manner using the SRA approach.[56] The committee designed the approach to fully comply with the *Evaluator Protocols* and the G*uidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches.*

The committee's goal was to develop a consistent set of survey questions with specific fill-in areas where the questions could be customized for the program being evaluated while maintaining consistency across the portfolio. The questions proposed by the committee for the standardized battery start on page 962 of the document. Customized surveys by program are also contained in the appendix.

The residential/small commercial NTGR algorithm derived four separate measurements of free-ridership:

1. The first consisted of responses to a series of yes/no questions that measured the impact of the program on the quantity, efficiency, and timing of the purchase.
2. The second consisted of a 0 to 10 scale that asked the likelihood that the respondent would have purchased the same exact high-efficiency measure in the absence of the program.
3. The third measurement combined responses to the quantity and timing questions with responses on a 0 to 10 scale to a question that asked the respondent's agreement with the statement that, in the absence of the program, they would have paid the additional rebate amount to buy the high-efficiency equipment on their own.
4. The final measurement combined responses to the quantity and timing questions with responses on a 0 to 10 scale to a question that asked the respondent's agreement with the statement that the program was a critical factor in their decision to purchase the high-efficiency equipment.

These questions covered all the requirements provided in the *Guidelines*, such as multiple questions, efficiency level, likelihood of adoption, timing and quantity, and consistency checks.

When responses are inconsistent among the four measurements, the guidelines prescribe that an analyst review the responses to open-ended questions that ask for clarification of the

[56] The Residential Retrofit High Impact Measure Evaluation Report and Appendices, (February 8, 2010, Prepared for the California Public Utilities Commission Energy Division) contains the questions used in the residential battery. http://www.calmac.org/publications/FinalResidentialRetroEvaluationReport_11.pdf http://www.calmac.org/publications/FinalResidentialRetroEvaluationAppendices.pdf

inconsistency and recode the four measurements as needed. The four separate measurements are then averaged to develop a final free-ridership estimate at the measure level. Prior to finalizing the NTGR algorithm, the committee conducted iterative testing with a partial data set. This testing contributed to the reliability of the algorithm and its computer coding.

The battery developed by the Residential NTG Committee is long and asks a number of very detailed questions. A more recent impact evaluation conducted for the California Public Utilities Commission for the Home Upgrade Program[57] used a more streamlined battery of questions to assess the extent of free-ridership in the program. The participant survey followed CPUC guidelines to assess free-ridership based on SR responses, considering respondent fatigue, complexity, timing, and budget constraints. The survey was finalized based on input from the PAs.

The Home Upgrade Program study defined full free riders as those who would have installed exactly the same measure at the same quantity (Q), efficiency (E), and time (T) in the absence of the program. The survey captured both full and partial free-riders who would have undertaken/installed the measure(s), but of lesser quantity, and/or at lesser efficiency, or a different time. Respondents selected one of two options when they began the survey: whether they (1) considered the project as one decision or (2) considered each measure installed as a separate buying decision. If they considered the project as one purchase decision, they received a short-form battery that applied to the entire project. If they considered each measure installed as a separate buying decision, they were given a long-form battery for each measure they reported installing. Those who received the short form battery were not asked the quantity question since they considered the whole project as one single decision.

The questions and algorithms used for the short-form and long-form batteries are contained in Appendix F of the final report and are shown below. Prior to asking the free-ridership questions, the survey included a number of project details-warm up questions.

Short-Form Scoring

The scoring process is summarized in Table 14 and Table 15. If the overall respondent level free-ridership is not determined by the first question on likelihood to implement the program, respondent level free-ridership score is the product of the timing and efficiency free-ridership scores. Responses in the scoring example included in the report are shaded light blue. The overall respondent level free-ridership in this example equals 0.25*0.5=0.125.

[57] Impact Evaluation Report, Home Upgrade Program – Residential Program Year 2017, prepared for the California Public Utilities Commission, April 29, 2019.
http://www.calmac.org/publications/CPUC_GroupA_Res_PY2017_HUP_toCALMAC.pdf

Table 14: Short-form Scoring

Without the program, how likely would you have been to undertake this project?	
Response option	**Overall respondent level free-ridership score**
Very likely	Next question on Timing
Somewhat likely	Next question on Timing
Somewhat unlikely	0.25
Very unlikely	0
Don't know	., Next question on Timing
Without the program, when would you have undertaken this project?	
Response option	**Timing free-ridership score**
At the same time or sooner	1
1 to 24 months later. Please specify the number of months = 18	1-(number of months/24) = 1-(18/24) =.25
More than 24 months later	0
Never	0
Don't know	.
Without the program, would you have installed (MEASURES) . . .?	
Response option	**Efficiency level free-ridership score**
That was the same or higher efficiency as what you installed	1
Above minimum standards/building code but lower efficiency than what you installed	0.5
Minimum standard/building code	0
Don't know	.

Long-form Scoring

If the overall respondent-level free-ridership is not determined by the first question on likelihood to implement the program, the respondent-level free-ridership score is the product of the quantity, efficiency, and timing free-ridership scores. If any of the three free-ridership scores in the product are zero, then the measure-level free-ridership is zero and the program gets full credit for the participant installing the measure. Measure-level free-ridership scores for a respondent are averaged to arrive at that respondent's overall free-ridership score.

Responses in the scoring example included in the final report are shaded light blue. The overall respondent level free-ridership in this example equals 0.25*0.5=0.125.

Table 15: Long-form Scoring

Without the program, would you say your likelihood on installing (SPECIFIC MEASURE) was?	
Response option	**Overall respondent level free-ridership score**
Very likely	Next question on Timing
Somewhat likely	Next question on Timing
Somewhat unlikely	0.25, next applicable measure
Very unlikely	0, next applicable measure
Don't know	., Next question on Timing
Without the program, when would you have installed (SPECIFIC MEASURE)?	
Response option	**Timing free-ridership score**
At the same time or sooner	1
1 to 24 months later. Please specify the number of months	1-(number of months/24)
More than 24 months later	0
Never (GO TO NEXT APPLICABLE MEASURE)	0
Don't know	.
Without the program, when would you have installed (SPECIFIC MEASURE)?	
Response option	**Timing free-ridership score**
At the same time or sooner	1
1 to 24 months later. Please specify the number of months	1-(number of months/24)
More than 24 months later	0
Never (GO TO NEXT APPLICABLE MEASURE)	0
Don't know	.
Insulation is rated with an "R-value", where the higher the R-value, the better the insulation's effectiveness. Without the program, would you have installed (INSULATION) with. . .?	
Response option	**Efficiency level free-ridership score**
Same or higher R value	1
Lower R value but above minimum standard/code	0.5
Minimum standard/building code	0
Would not have installed any insulation (GO TO NEXT APPLICABLE MEASURE)	0
Don't know	.
Without the program, would you have . . .?	
Response option	**Quantity level free-ridership score**
Covered the same area/square feet (200%)	1
Covered < 100% but more than 0% = 60%	1 – (response percent/100) = 1-.6=.4
Would not have installed insulation (0%)	0
Don't know	0

A.6 NEEP Decision-Framework for Determining Net Savings Approach – Supplemental Document #2 to Principles and Guidance, September 2014

This document summarizes three literature reviews and white papers on measuring net savings.[58] It highlights the different methods that can be used to estimate net savings under different conditions, such as for custom measures, measures with few and diverse participants, measures with large numbers of similar participants, and measures with upstream influence. It also rates the typical cost or complexity of each method as well as any special requirements. This document also groups the various approaches on a continuum identifying those techniques that are driven heavily by data compared to those techniques that are driven heavily by judgement. This document does not recommend specific survey questions or algorithms.

[58] NMR Group, Tetra Tech & KEMA 2011, "Massachusetts Program Administrators Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches (Final), July; Violette D. & Rathbun, P. 2014, "Chapter 17: Estimating Net Energy Savings: Common Practices," Uniform Methods Project, September; and Tetra Tech, KEMA & NMR 2011, "Massachusetts Program Administrators Cross-Cutting C&I Free-Ridership and Spillover Methodology Study Final Report, April.

Appendix B Alternative Approaches for Estimating NTG

While the methods for estimating net program impacts for Massachusetts' downstream C&I programs has been standardized since 2003, the methods for estimating these impacts for the residential programs have never been standardized. In 2011, the Special and Cross Cutting team developed a report with suggested approaches for consideration by the PAs for estimating NTG effects for residential programs.[59] Although each of the approaches contained in that report have the potential for identifying net savings, specific study designs and results need careful review to minimize or characterize the potential for inaccurate results.

The Special and Cross Cutting team developed a 2017 Net-to-Gross Methodology Research (TXC08) report,[60] which described the factors noted by other researchers that evaluators should consider regarding each potential NTG approach (NMR et al. 2011, the NEEP Decision-Framework for Determining Net Savings Approach,[61] and the UMP Net Savings chapter).[62] The approaches and the factors contained in that report are summarized in the remainder of this Appendix. For more detail on these approaches, the reader should refer to the three documents cited above.

B.1 Billing Data Analysis

- Billing analysis typically requires up to 12 months of post-implementation consumption data. Generally, billing analysis is only appropriate when participant whole-house or facility savings are substantial relative to total consumption and when there are large numbers of fairly homogenous participants (NMR et al., 2011, p. 4).
- Non-participant spillover savings can count against the program to the degree those savings are present in a comparison group. In other words, to the extent that program savings *leak* into the comparison group, they will be assumed to be part of general market trends and not attributed to the program as they should be.
- Billing analysis savings estimates are generally designed to provide aggregate program savings, possibly by subgroup, not individual participant net savings.

---

[59] Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches, Prepared for the Massachusetts Program Administrators by NMR Group, Inc. with contributions by Tetra Tech and KEMA, June 20, 2011. http://ma-eeac.org/wordpress/wp-content/uploads/Cross-Cutting-Net-to-Gross-Methodology-Study-for-Residential-Programs-Suggested-Approaches-Final-Report.pdf

[60] http://ma-eeac.org/wordpress/wp-content/uploads/Net-to-Gross-Methodology-Research.pdf

[61] Johnson Consulting Group (2014). Decision-Framework for Determining Net Savings Approach – Supplemental Document #2 to Principles and Guidance. Prepared for NEEP.

[62] Chapter 21: Estimating Net Savings –Common Practices: Methods for Determining Energy-Efficiency Savings for Specific Measures, Daniel M. Violette and Pamela Rathbun, 2017. Golden CO: National Renewable Energy Laboratory. NREL/SR-7A40-68578. https://www.nrel.gov/docs/fy17osti/68578.pdf

- This approach depends heavily on the comparability of the comparison group. Study design should consider whether the comparison group represents net or gross savings. For example, if more recent participants are used as the comparison group, the savings may only represent gross savings.
- The net savings from billing analysis includes, without separating estimate, participant non-free rider savings, participant spillover, non-participant spillover, interactive effects, and rebound effects.
- Self-selection can distort the net savings estimates, to the extent the tendency to join the program or not is associated with a tendency to change consumption in other ways. Some analytic methods exist to limit these effects, but outside of randomized treatment control designs the effect cannot be fully eliminated.
- Major decision criteria are as follows:
  - Are data available from an appropriate comparison group?
  - Are non-participant SO savings expected to be significant?
  - Is the approach designed to minimize self-selection bias?

B.2 Market Sales Data Analyses

- Market sales data analyses measures the total net effect of the program, including both FR and participant and non-participant *like* SO. The most common approach is a cross-sectional comparison area method in which post-program data are compared with data from a non-program comparison area (or multiple comparison areas) for the same point in time.
- The results inform a net to gross ratio, which is applied to gross program savings.
- Market sales data analyses are dependent on the availability and quality of sales and shipment data in the area of interest and comparison areas
- Comparison areas may be subject to influence by the program or other programs and should be carefully considered before selecting.
- Impacts are subject to potential bias if market data is limited and without careful selection of a comparison area.
- Major decision criteria: there are a number of important factors to consider when determining if a market sales data analysis is appropriate:
  - Does an appropriate comparison area exist?
  - Are the market data available and complete?
  - What are the features of the program and should a market impact outside of program-tracked project be expected to occur?
  - Are the market measures appropriate for upstream analysis (i.e., lighting or appliance programs)?

- How will this analysis be supplemented with additional approaches to provide context for these findings?

B.3 Pricing and Elasticity Analysis

- Many of these approaches rely on customers to identify their intentions, based on actual or stated preferences, compared to hypothetical situations.
- In stated preference models, there is a potential difference between what customers say their purchase intentions are in a hypothetical situation compared to what they actually do. Revealed preference approaches rely on actual purchase activity.
- Pricing and elasticity models that solely use sales program data and market pricing require adequate variability in prices to identify effects.
- The results are net of FR, with no measurement of SO. They inform a net to gross ratio, which is applied to gross program savings
- Major decision criteria are as follows:
  - How will a survey approach account for the "halo" or "Hawthorne"[63] effects?
  - Is there adequate pricing variability to rely solely on sales and pricing data?
  - How will this analysis be supplemented with additional approaches to provide context for these finding?
  - Are customers likely to be experiencing survey fatigue?

B.4 Self-Reported Counterfactual

- Despite its drawbacks, this approach is typically the most cost-effective, transparent, and flexible method for estimating NTG. Surveys may target different types of respondents including program participants, program non-participants, and market actors.
- End-user SR counterfactual is challenging for upstream programs, both because identifying customers who received program benefits is challenging, and because the program tends to be transparent to the customer. When customers can be identified, it is possible to ask them what they would have done at other price points or if the product had not been available in a particular store. Manufacturers or retailers may provide a better source of information for upstream program.
- The results inform a net to gross ratio, which is applied to gross program savings.
- Participant surveys only capture a subset of market effects or program attribution.
- SR assessments of program influence are subject to a number of sources of unknown bias. These include the following:

[63] The **Hawthorne Effect** is the process where human subjects of an experiment change their behavior because they are being studied.

- Consumers inability to know what would have done in hypothetical counterfactual situation, the potential for rationalizing past actions or give socially desirable answers, or a general failure to recognize the influence of a program
- Potential arbitrariness of scoring methods in algorithms

- Major decision criteria are as follows:
  - Is the program or measure type suitable for this approach?
  - What is the potential for response bias and nonresponse bias?
  - How will this analysis be supplemented with additional approaches to provide context for these findings?
  - Are customers likely to be experiencing survey fatigue?

B.5 Structured Expert Judgment

- A particularly useful role for structured expert judging is to develop a consensus estimate to consolidate results from multiple estimation methods or complex data. The Delphi process is the most widely known technique.
- The approach relies on the availability of high-quality data to inform the panel, leading to reasonable estimates of program attribution.
- Sampling-based calculations of confidence and precision are not available.
- Major decision criteria are as follows:
  - Is the approach designed to minimize bias among the panel?
  - Is the program or measure type suitable for this approach?
  - How will this analysis be supplemented with additional approaches to provide context for these findings?

B.6 Historical Tracing

- This method involves the careful reconstruction of events leading to the outcome of interest to develop a “weight of evidence” conclusion regarding the specific influence or role of the program in question on the outcome. This approach suggests that if the predicted steps between an activity and an outcome can be confirmed in implementation, this lends a strong argument for causality. Because this method draws from multiple information sources, it is difficult to determine the magnitude of the effects. Rather, it provides evidence to a Delphi panel that they could consider quantifying a net savings or net to gross ratio estimate.
- Best suited to attribution analysis of major events, such as adoption of new building codes or policies, and is not typically applicable to energy-efficiency programs.
- It may be difficult to translate the influence factors into estimates of impacts without additional modeling.

- Evaluator cannot calculate sample-based statistical confidence and precision levels for the estimate.
- Major decision criteria are as follows:
  - Is the historical data available to minimize bias?
  - Is the program or measure type suitable for this approach?
  - How will this analysis be supplemented with additional approaches to provide context for these findings?

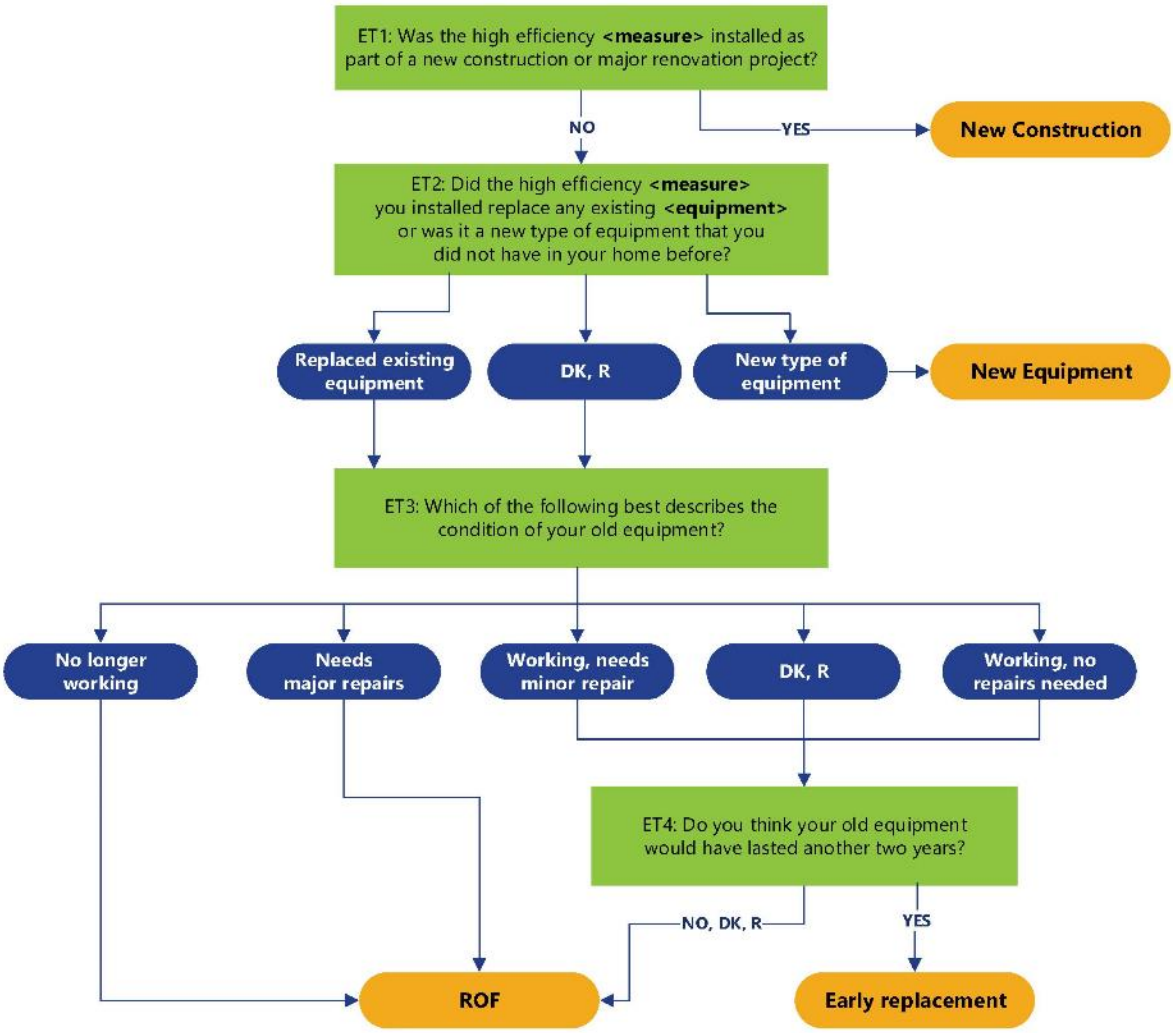
B.7 TOP-DOWN ECONOMETRIC MODELING

- Top-down evaluations use macro data on energy consumption in a model that relates changes in energy consumption to a measure of energy-efficiency program activity, such as program expenditures.
- This approach is subject to the same potential bias as any cross-sectional and/or longitudinal regression analysis due to the influence of confounding variables that are not accounted for in the model. Use of comparison areas can mitigate some of this effect but it can also exacerbate it by introducing additional unaccounted sources for variation.
- Methods are not fully developed at utility, state or regional levels.
- Relies on high-quality consumption data and data on energy-efficiency program efforts over time.
- Top-down methods provide savings at the portfolio or sector level and do not provide savings at the measure, technology or program level.
- The methods provide average savings factors over several years included in the study period, such as savings per dollar of spending or per bottom-up estimated MWh saved. They do not provide separate factors of this type by program year.
- The net savings from the top-down analysis include, but do not separately estimate participant non-FR savings, PSO, NPSO, interactive effects, and rebound effects.
- Depending on the level of geographic aggregation used, cross-unit SO can count incorrectly against the net savings.
- Depending on the level of geographic aggregation used, participant self-selection can distort the net savings estimates if not accounted for analytically.
- Does not provide information on how to improve program design and implementation and shows savings only after a sufficient number of years have passed with continuous program portfolio operations.
- Major decision criteria:
  - Is there adequate available data to understand program expenditures and energy consumption?
  - Can lag effects be modeled with the available time span of available data?

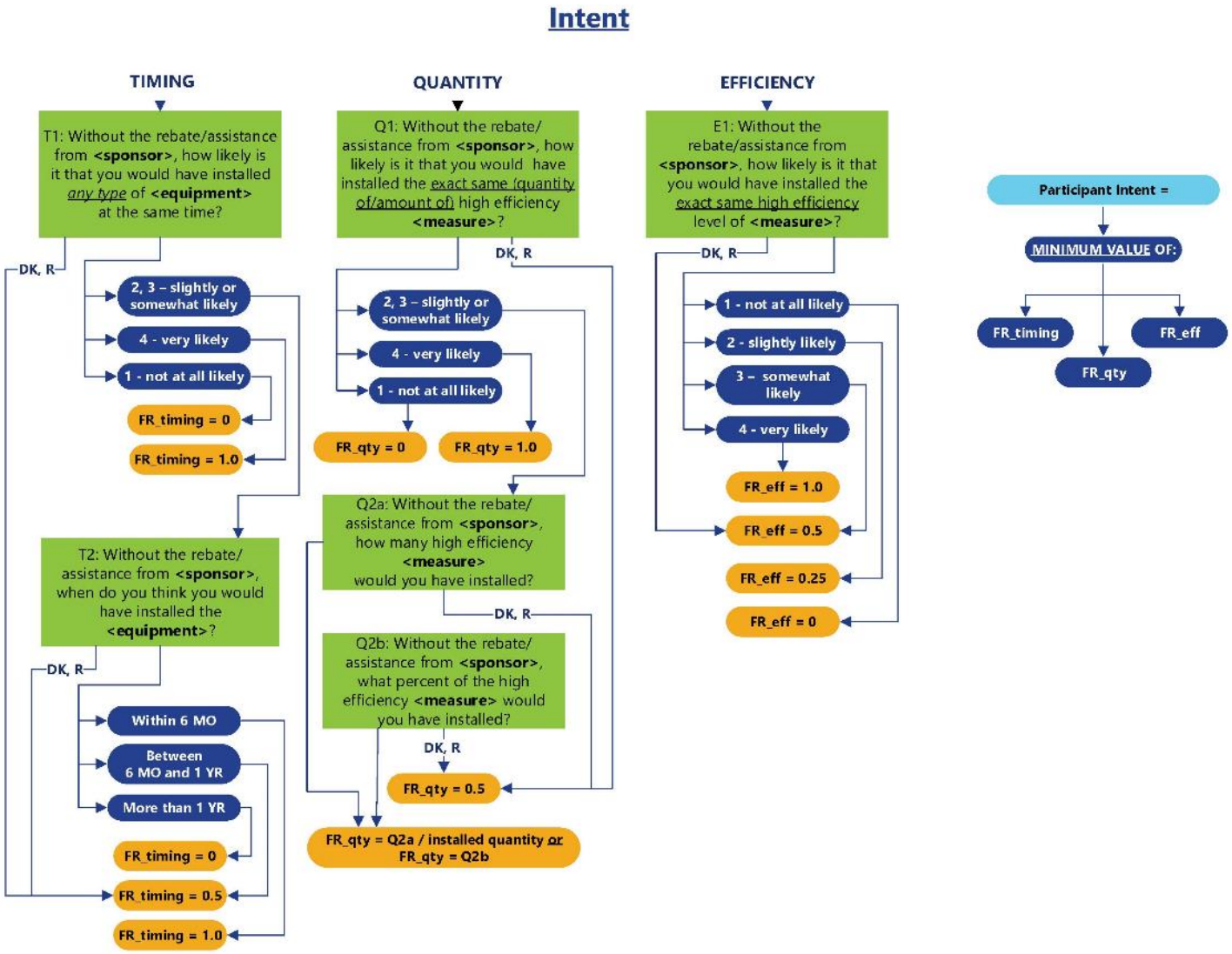
- How will this analysis be supplemented with additional approaches, e.g., program-specific process and/or impact evaluations, to provide context for these findings?

Appendix C Algorithms for the Recommended Residential NTG SR Survey Questions

C.1 Event Type Algorithm

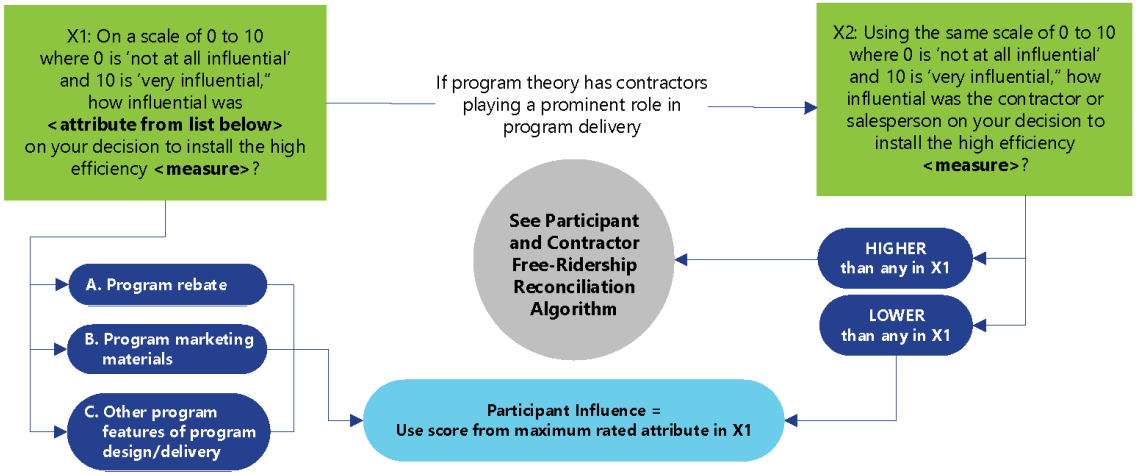


C.2 PARTICIPANT FREE-RIDERSHIP ALGORITHM



Participant Free-Ridership Algorithm

Influence



Calculation of the Influence Score

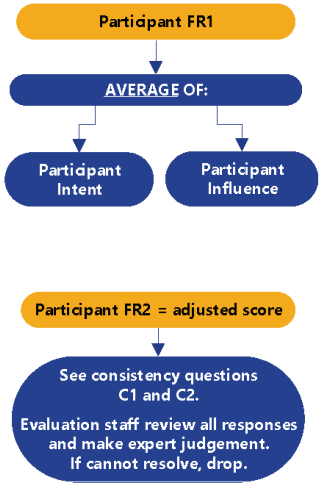
Rating	Score
Not at all influential = 0	1
1	0.9
2	0.8
3	0.7
4	0.6
5	0.5
6	0.4
7	0.3
8	0.2
9	0.1
Very influential = 10	0

Consistency

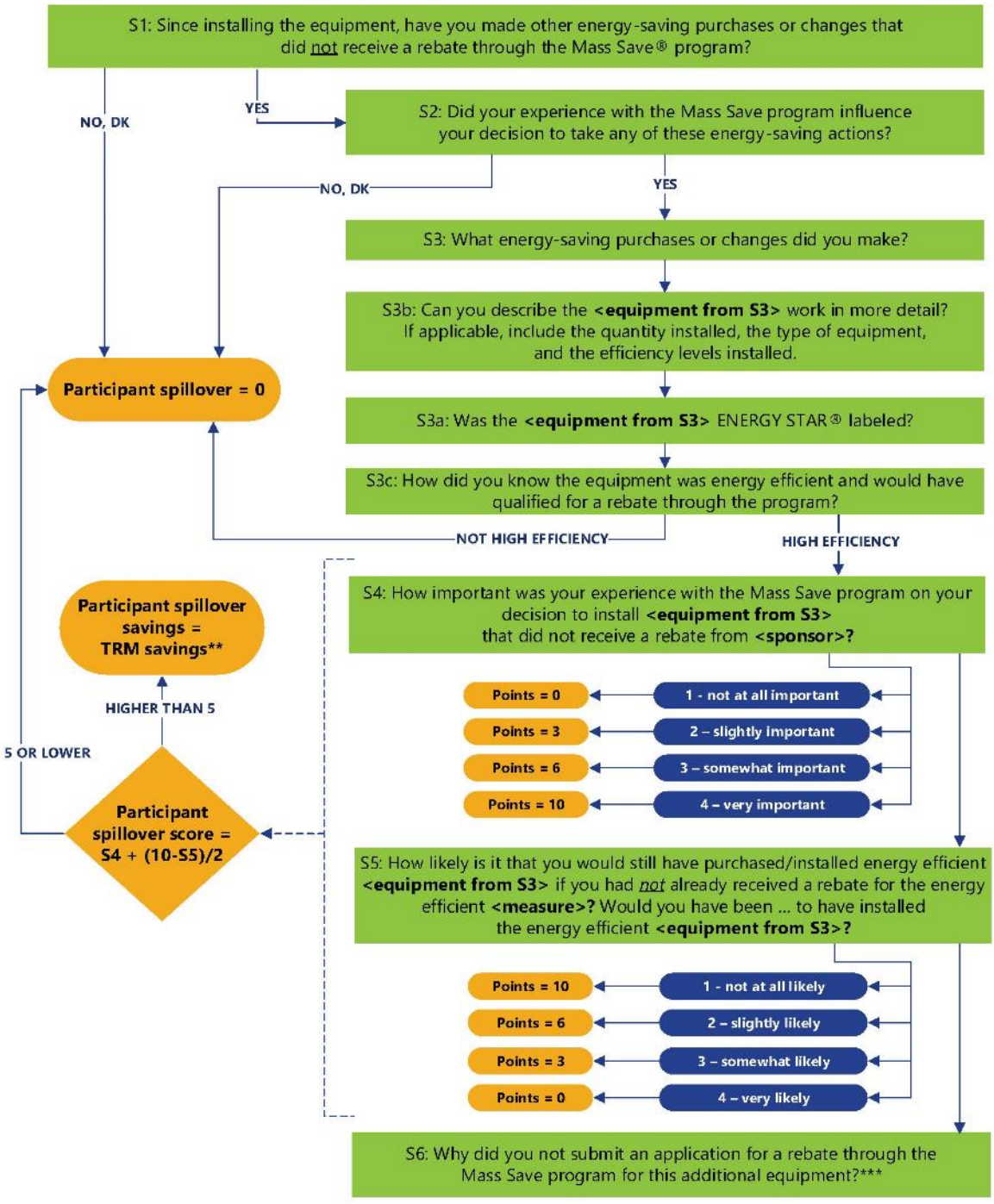
(Ask if participant intent and participant influence are inconsistent)

C1: Did you make the decision to install the high efficiency <measure> before or after you learned of the rebate?

C2: Please describe the influence that the <program rebate/assistance> had on your decision to install the high efficiency <measure>?



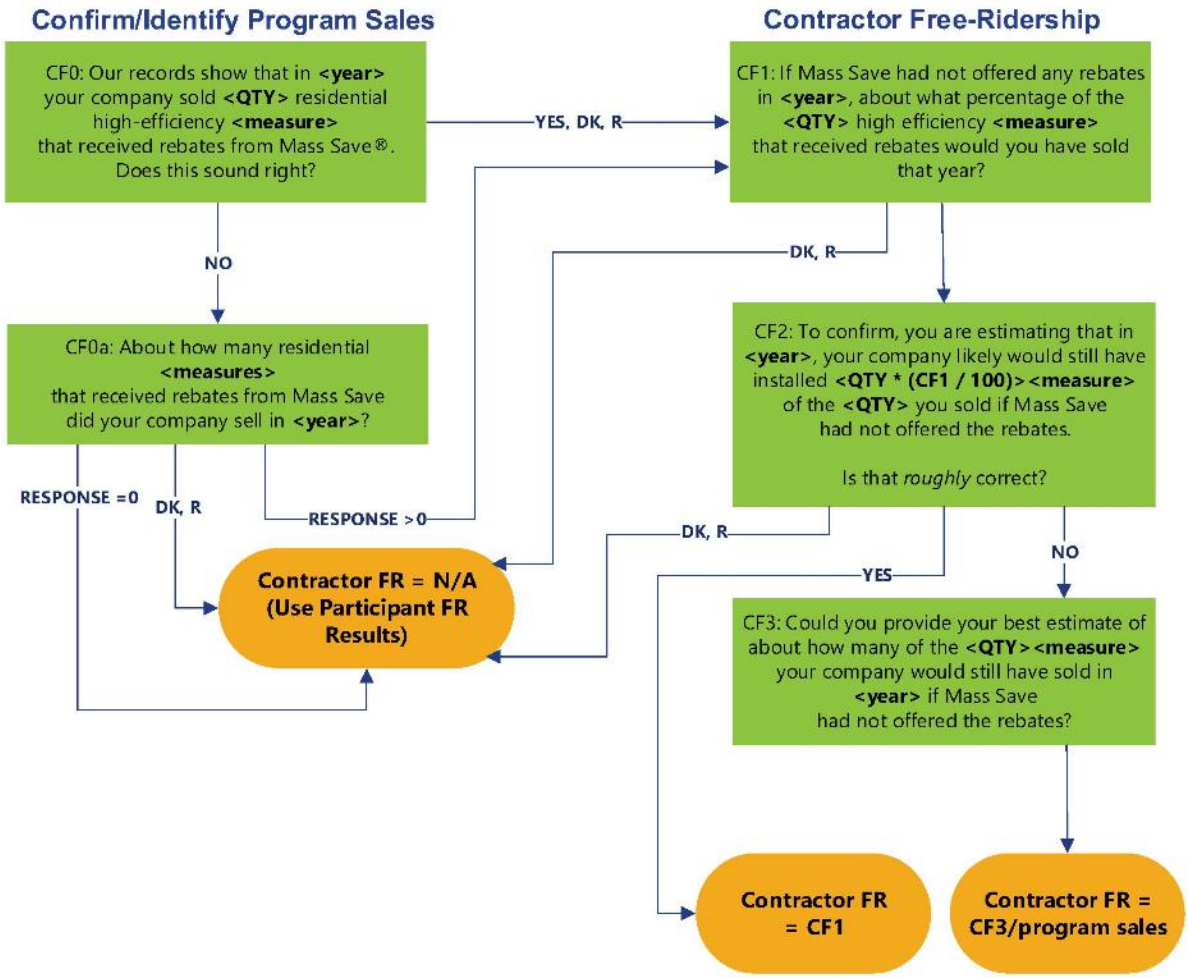
C.3 PARTICIPANT SPILLOVER ALGORITHM



* Efficiency specifications use TRM specifications for measures eligible for PA programs. For measures not offered by the PAs, the team will determine if the measures were higher than federal standards. Customers were explicitly asked if appliances were ENERGY STAR-qualified.
** Secondary sources will be used to estimate savings for non-like measures if they are not available in the MA TRM. Spillover rates are equal to the sum of the savings associated with the spillover measures divided by the sum of the savings associated with respondents participating measures.
*** Consistency check.

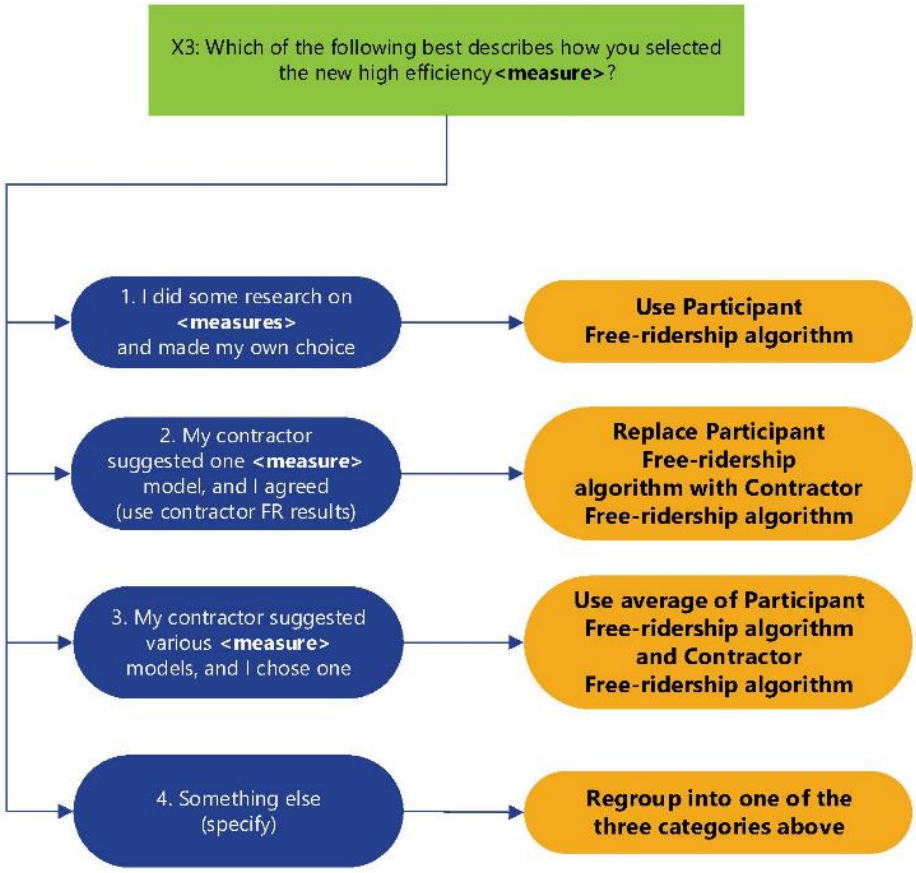
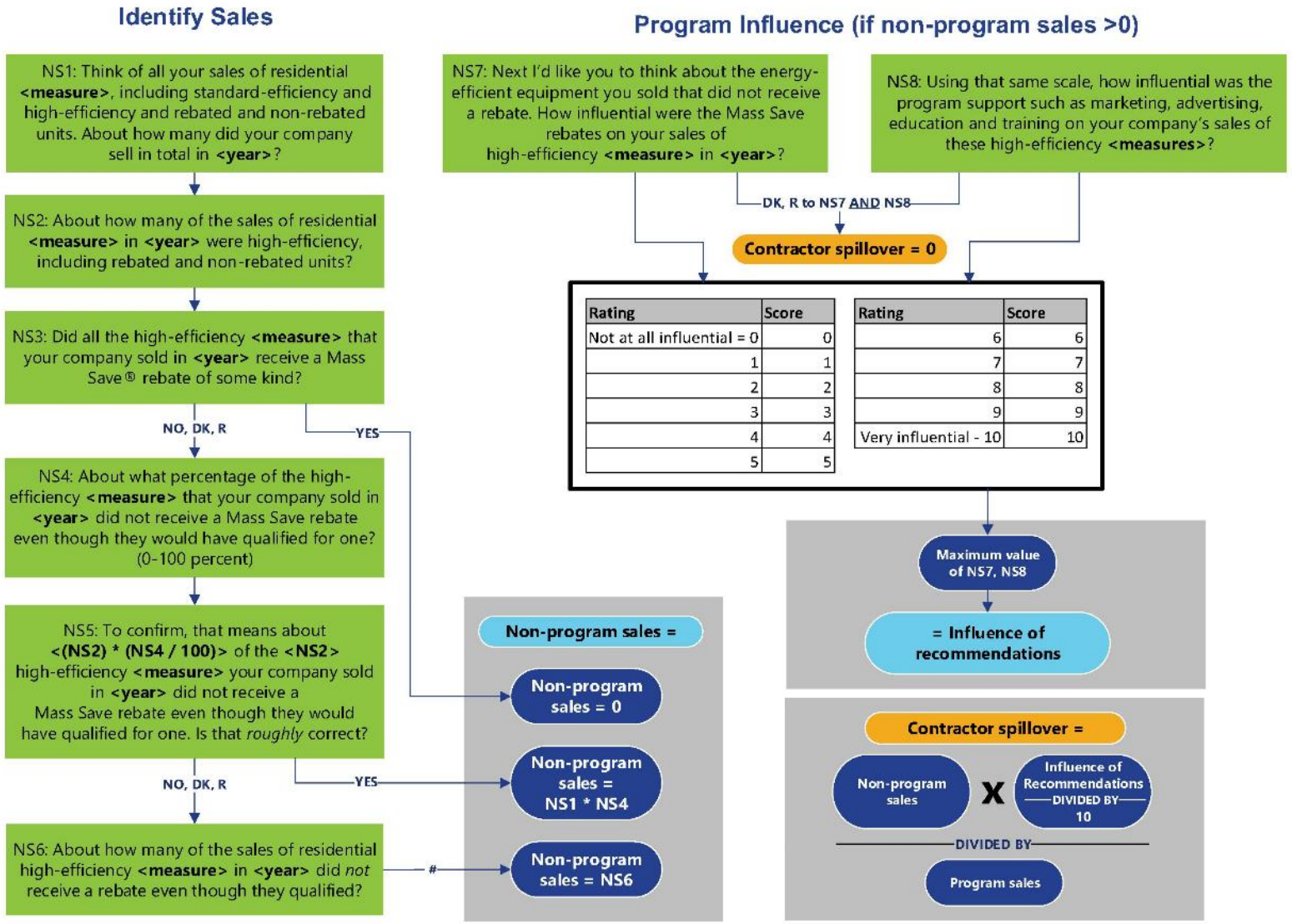
C.4 CONTRACTOR FREE-RIDERSHIP ALGORITHM

If program theory shows that the contractor/salesperson plays a prominent role in delivering and promoting the program.



C.5 PARTICIPANT AND CONTRACTOR FREE-RIDERSHIP RECONCILIATION

When the contractor was rated the most influential program element in Participating FR and the program theory shows that the contractor pays a prominent role in the program's delivery, use this algorithm.



C.6 PARTICIPATING CONTRACTOR SPILLOVER ALGORITHM