SEE Action

STATE & LOCAL ENERGY EFFICIENCY ACTION NETWORK

Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs:

Issues and Recommendations

SEE Action

STATE & LOCAL ENERGY EFFICIENCY ACTION NETWORK

Outline of Presentation

- Report Objective, Scope, and Outline
- Key Takeaways
- Technical Recommendations
- Questions

Outline of Presentation

- Report Objective, Scope, and Outline
- Key Takeaways
- Technical Recommendations
- Questions

Objective of Report

- This report provides guidance and recommendations on methodologies that can be used for estimating energy savings impacts resulting from residential behavior-based efficiency programs
  - Ensure that estimated program savings impacts are valid

- Two primary target audiences:
  - Senior managers responsible for overseeing and reviewing efficiency program designs and evaluations
    → Executive Summary
  - Practitioners, evaluation professionals, and staff responsible for reviewing efficiency program designs and evaluations
    → Main Technical Report

Scope of Report

- This report provides guidance and recommendations on methodologies that can be used for estimating **energy savings** impacts resulting from **residential behavior-based** efficiency programs

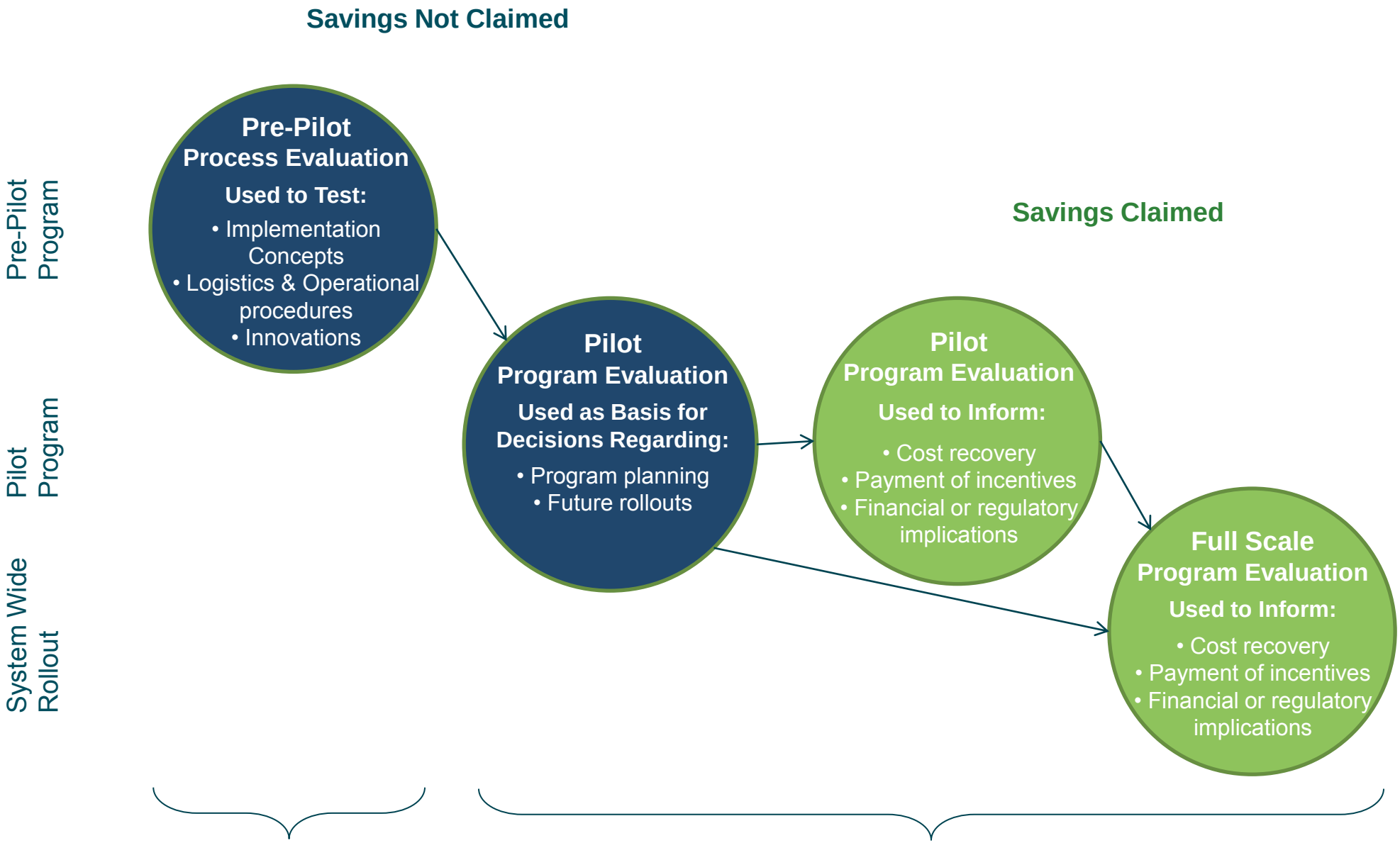
Behavior-based energy efficiency programs are those that utilize strategies intended to affect consumer energy use behaviors in order to achieve energy and/or peak demand savings. Programs typically include outreach, education, competition, rewards, benchmarking and/or feedback elements.

Outcome of interest is **energy savings defined at the household level**

- Not peak demand savings, time based tariff programs, market transformation, participation levels, # of rebates, etc.

SEE Action
STATE & LOCAL ENERGY EFFICIENCY ACTION NETWORK

Scope: Typical Program Life Cycle



Outline of the Report

- Executive Summary
- Concepts and Issues in Estimation of Energy Savings
  - Explanation of RCTs
  - Explanation of quasi-experimental methods
- Technical Recommendations
  - 14 evaluation design and analysis issues (details in following slides)
  - Based on consensus of researchers in many different fields and environments
  - Vetted by ~50 reviewers: technical, program administrators, academics, regulatory agencies, industry stakeholders
- Four Examples of Behavior Based Programs
  - 3 RCTs, 1 Quasi-experimental

Outline of Presentation

- Project Timeline and Process
- Report Objective, Scope, and Outline
- Key Takeaways
- Technical Recommendations
- Questions

Key Takeaways

- **Key recommendation: use a Randomized Controlled Trial (RCT) to evaluate behavior-based programs**
- How did we reach this conclusion? Tradeoffs:
  - Behavior-based program impact estimates are **uncertain** / risky; relatively new; savings are relatively small
  - RCT method is robust, and offers a **high degree of confidence** in the validity of program impact estimates
  - RCT method is potentially less practical than quasi-experimental methods…but quasi-experimental methods are not as robust and are potentially biased
  - If RCT not feasible, quasi-experiment acceptable
- **Key future research: create a model that accurately predicts program savings in future years, new populations**
  - Move away from RCTs; expand program system wide

Outline of Presentation

- Project Timeline and Process
- Report Objective, Scope, and Outline
- Key Takeaways
- Technical Recommendations
- Questions

Technical Recommendations

Technical recommendations solve two areas of concern:

- **Internal Validity** - methods to ensure that estimated program savings impacts are valid for a given program participant population in a given time frame
- **External Validity** - methods for applying estimated program savings impacts to new populations and future years

Technical Recommendations

- **Internal Validity** - methods to ensure that estimated program savings impacts are valid for a given program participant population in a given time frame
  - Evaluation design
  - Length of study, baseline period
  - Avoiding potential conflicts of interest
  - Analysis estimation methods
  - Statistical significance
  - Double counting issues
- **External Validity** - methods for applying estimated program savings impacts to new populations and future years

Evaluation Design

Method in which a control group (customers not in the program) is constructed and compared to the treatment group (customers in the program) in order to estimate the program savings impact

Randomized Controlled Trial results in unbiased estimates of savings

Regression Discontinuity results in estimates of savings that are likely to be unbiased if done correctly

Variation in Adoption with a Test of Assumptions could result in biased estimates of savings[53]

Propensity Score Matching could result in biased estimates of savings[54]

Non-Propensity Score Matching could result in biased estimates of savings

Pre-Post Comparison could result in very biased estimates of savings

Length of Baseline Period

Relatively longer study periods and baseline data periods are likely to lead to greater precision of the estimated program impact

If RCT:	If Quasi-Experimental:	
★★★★★	★★★★★	Twelve Months or More of Historical Data Collected[57]
★★★★☆	Not Advisable	Less Than a Complete Twelve Months of Historical Data Collected
★★★★☆	Not Advisable	No Historical Data Collected

Avoiding Potential Conflicts of Interest

Evaluations should be managed in a way that produces the least potential for a conflict of interest to arise regarding the validity of savings estimates

An independent, third-party evaluator transparently defines and implements:

- Program evaluation
- Assignment of households to control and treatment groups
- Data selection and cleaning, including identification and treatment of missing values, outliers, and account closures, and the normalization of billing cycle days

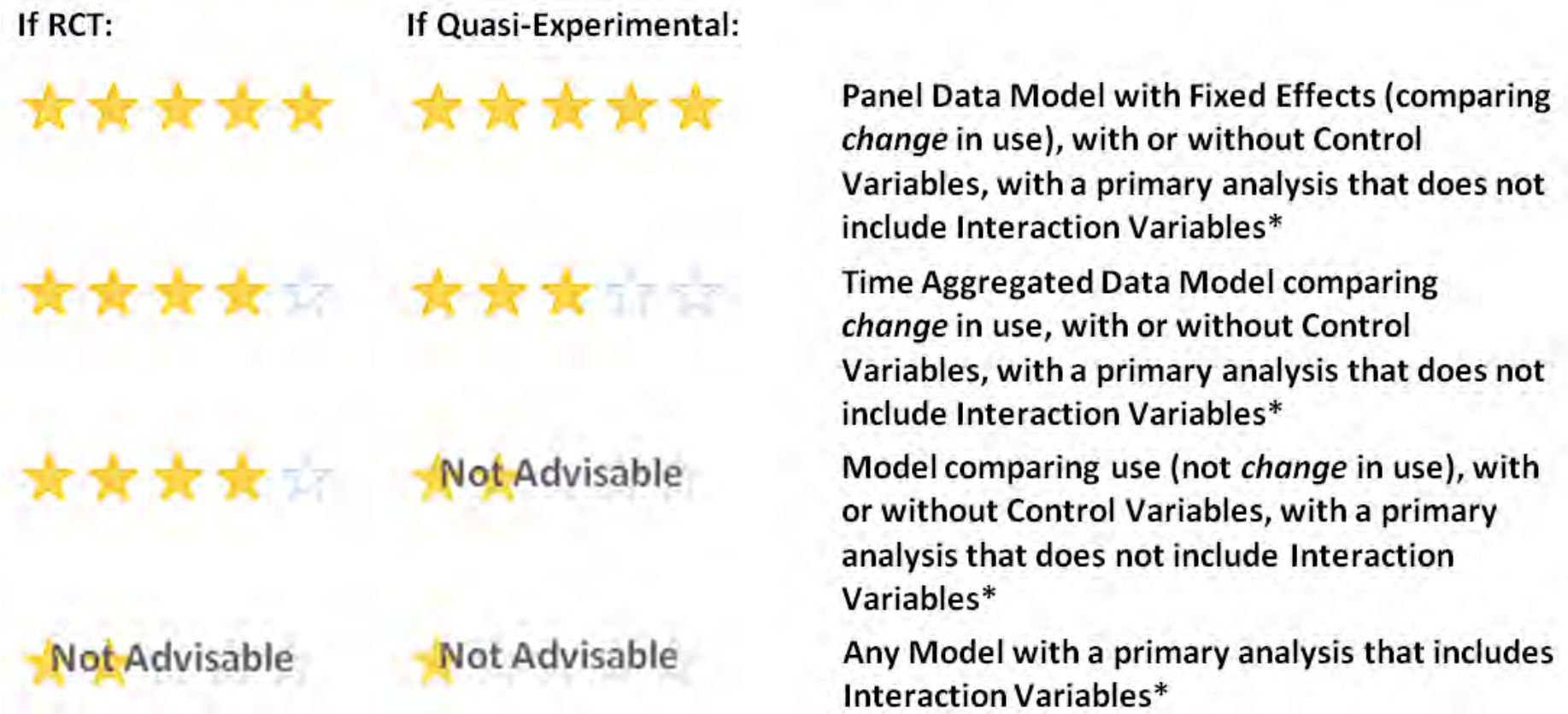
Program implementer or sponsor implements any of the above activities

Analysis Model Specification Options

Analysis model is the set of algorithms used to estimate energy savings through econometric techniques (e.g., regression analysis)

3 specification options:

- Change in energy usage vs. energy usage
- Use panel data (many data points over time) vs. aggregated data
- Include interaction variables vs. no interaction variables

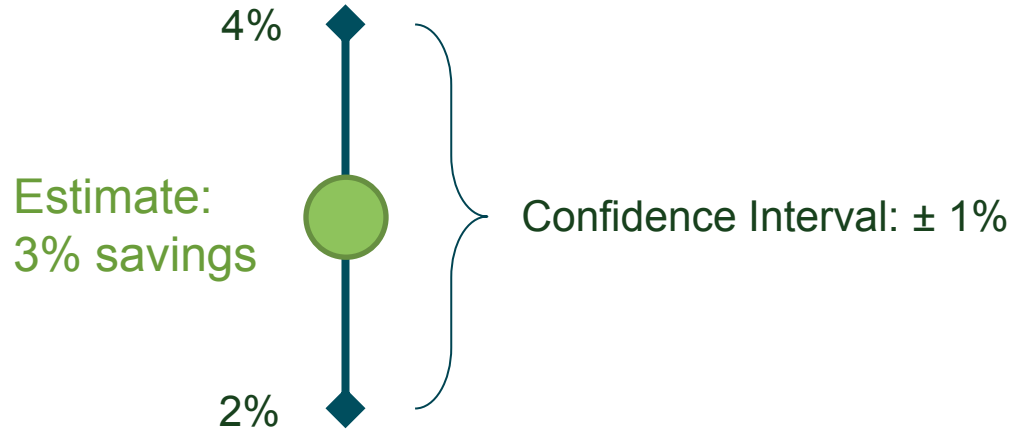


Statistical Significance

- **Problem**: how to ensure that the estimate of program impact savings is precise enough, not risky
- **Recommendation**:
  - Define null hypothesis (the required threshold, e.g., cost effectiveness)
  - Estimate considered acceptable if statistically significant at 5% (i.e., 95% confidence)
  - 5% statistical significance *NOT* the same as 95/5

Statistical Significance: Example

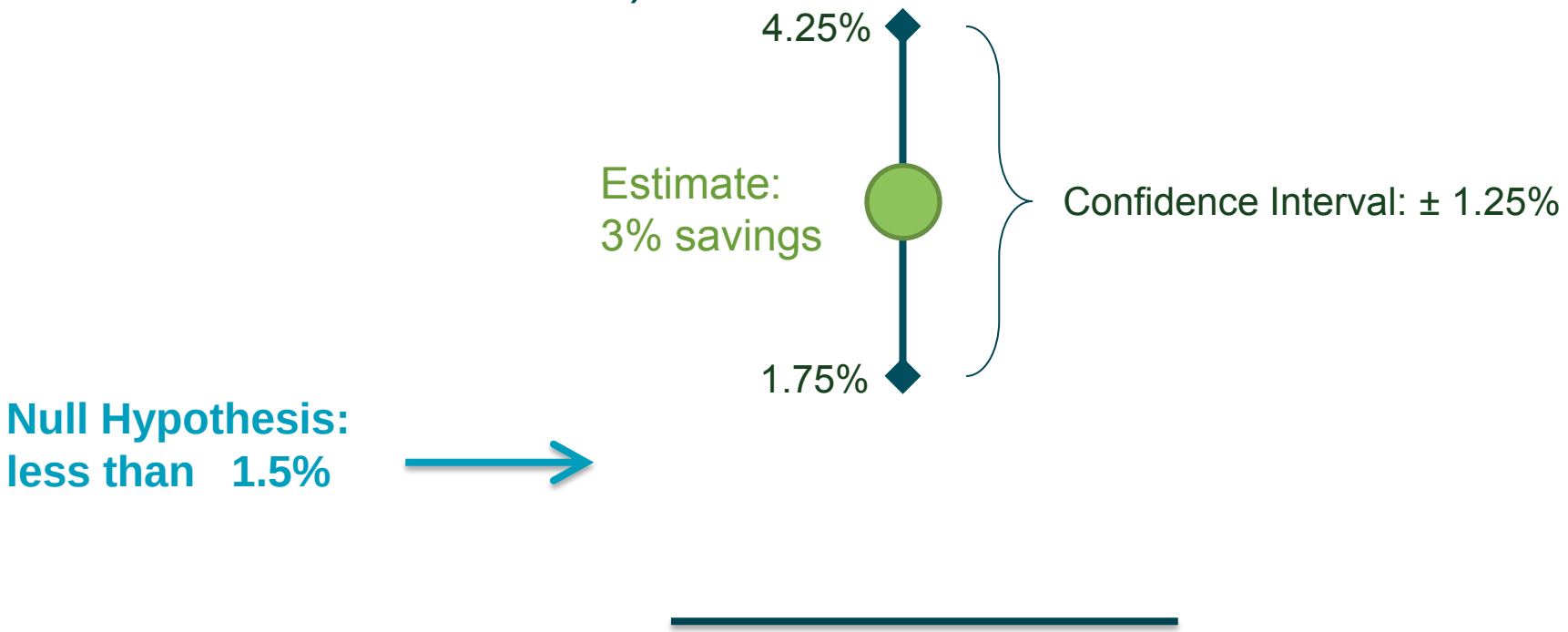
- **Example**: test whether program results in *positive* (i.e., greater than zero) energy savings
  - Null hypothesis = zero
  - 5% statistical significance (i.e., confidence interval does not include zero):



→ Conclude: program *does* result in energy savings

Statistical Significance: Example

- **Example**: test whether program results in *cost effective* energy savings (e.g., cost effectiveness threshold = 1.5%)
  - Null hypothesis: program is not cost effective, savings are less than 1.5%
  - 5% statistical significance (lower bound of confidence interval is above 1.5%):



→Conclude: program *is* cost effective

Statistical Significance: Example

- **Example**: test whether program results in *cost effective* energy savings (e.g., cost effectiveness threshold = 2.5%)
  - Null hypothesis: program is not cost effective, savings are less than 2.5%
  - *Not* 5% statistically significant (lower bound of confidence interval is *below* 2.5%):

→Conclude: program *may or may not be* cost effective (too risky to say for sure)

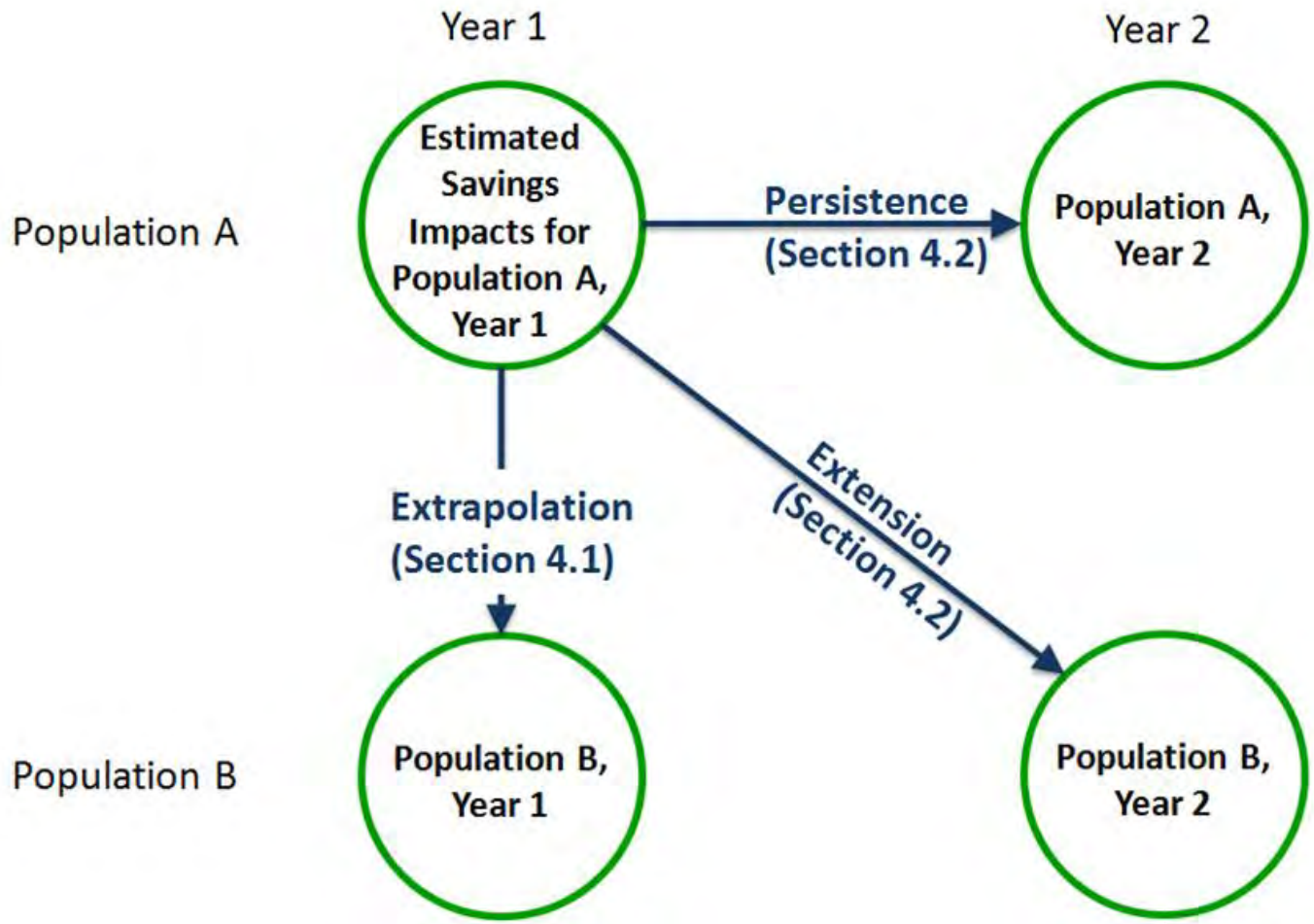
Accounting for Potential Double Counting of Savings

- **Problem**: the same savings may be claimed by two programs (e.g., a behavioral program & appliance rebate program both claim savings from appliances)
- **Recommendation:** estimate this "double counted savings" to the extent possible by comparing control to treatment group
  - For programs that can be *tracked* (e.g. installation of insulation by a contractor), rigorously estimate double counted savings;
  - For programs that cannot be tracked (e.g., upstream CFL rebates), attempt to estimate double counted savings;
  - The measurement period (e.g., accounting for seasonal load impacts), and the effective useful lifetime of installed measures (when lifetime savings are reported) are taken into account; and
  - Program costs are appropriately allocated along with double counted saving

Technical Recommendations

- **Internal Validity** - methods to ensure that estimated program savings impacts are valid for a given program participant population in a given time frame
- **External Validity** - methods for applying estimated program savings impacts to new populations and future years

External Validity



1. Extrapolate to new population (in the same year)
2. Persistence – extrapolate to future years (with the same population)
3. Extend – extrapolate to new populations in future years

Applicability of Impact Estimates from One Population to Another (in same year)

The applicability of program impact estimates from one population (population A) to another population (population B), depends on the similarity of population A to population B

Population A is **very similar** to population B

(population A is a random sample of larger population A+B)

Population A is **different** than population B

(e.g., population A has higher energy usage than B)

Results from A **are likely applicable** to large population A+B

Results from A have **limited applicability** to larger population A+B

Results from A **are not applicable** to larger population A+B

Persistence of Savings

- **Problem:** If the program continues over several years, does energy savings from behavior-based programs continue over time?
- **Recommendation:**
  - A control group be maintained for every year in which program impacts are estimated
  - Evaluation is done each year initially and every few years after it has been running for several years

Applicability of Impact Estimates to a New Population in Future Years

- **Recommendation: If the program is expanded to new program participant populations,** a control group that is representative of all of the different participating populations should be created and maintained for every population in the expanded program for every year in which program energy savings estimates are calculated.

Recommendations for the Future: Predictive Models

- The transition towards system wide behavior-based EE programs will result in significant pressure to develop a rigorous predictive model
- What will it take for a predictive savings model to be credible for claiming savings?
  - Many years of past data
  - Predicted years are similar to past years
  - Model is rigorously validated
  - Estimate is risk-adjusted: claimed savings = lower bound of the 95% confidence interval

Outline of Presentation

- Project Timeline and Process
- Report Objective, Scope, and Outline
- Key Takeaways
- Technical Recommendations
- **Questions?**

Mike Li: Michael.Li@hq.doe.gov

Annika Todd: atodd@lbl.gov

SEE Action

STATE & LOCAL ENERGY EFFICIENCY ACTION NETWORK

Additional Technical Recommendations

Excluding Data from Households that Opt-out or Close Accounts

Data cleaning: which households to exclude

★★★★★ Only data from households that closed accounts are excluded*; households that opt-out of the treatment or control group are included in the analysis (although the program impact estimate may be transformed to represent the impact for households that did not opt-out, as long as it is transparently indicated).

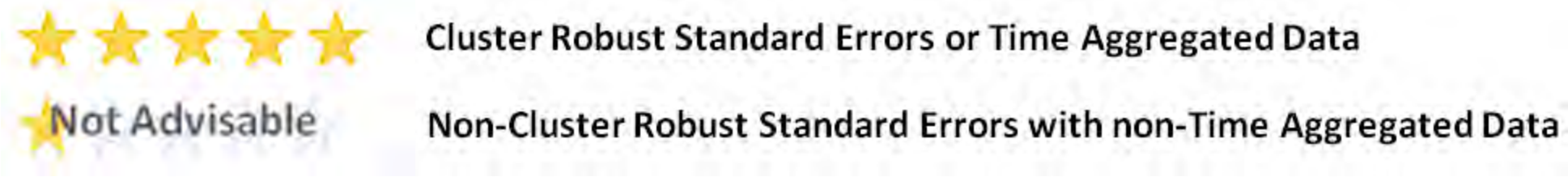
★ Not Advisable — Data from households that closed their accounts are included*

★ Not Advisable — Households that opt-out are excluded from the analysis

*If there is a compelling reason to include households that closed their accounts and the analysis is undertaken correctly to deal with unbalanced data sets, then it may be advisable.

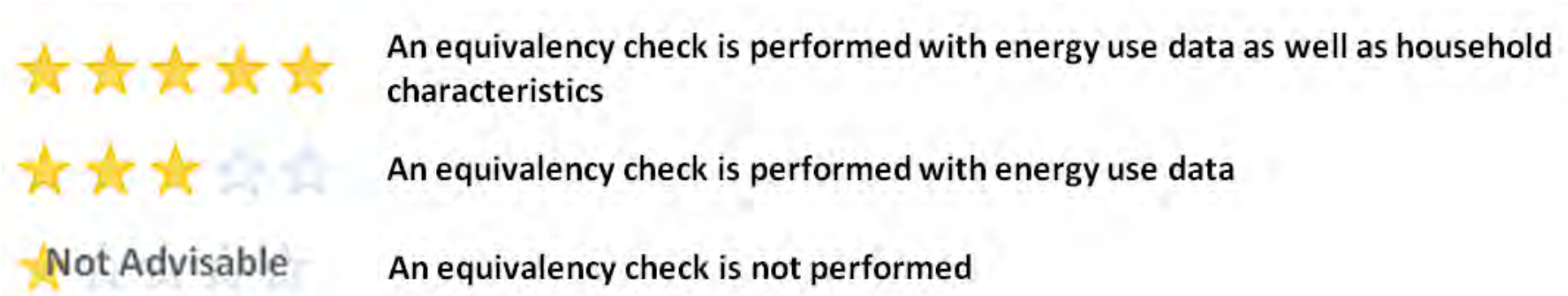
Cluster Robust Standard Errors

Ensure that the standard errors are robust and account for clustering



Equivalency Check

Validate that the control and treatment group are equivalent



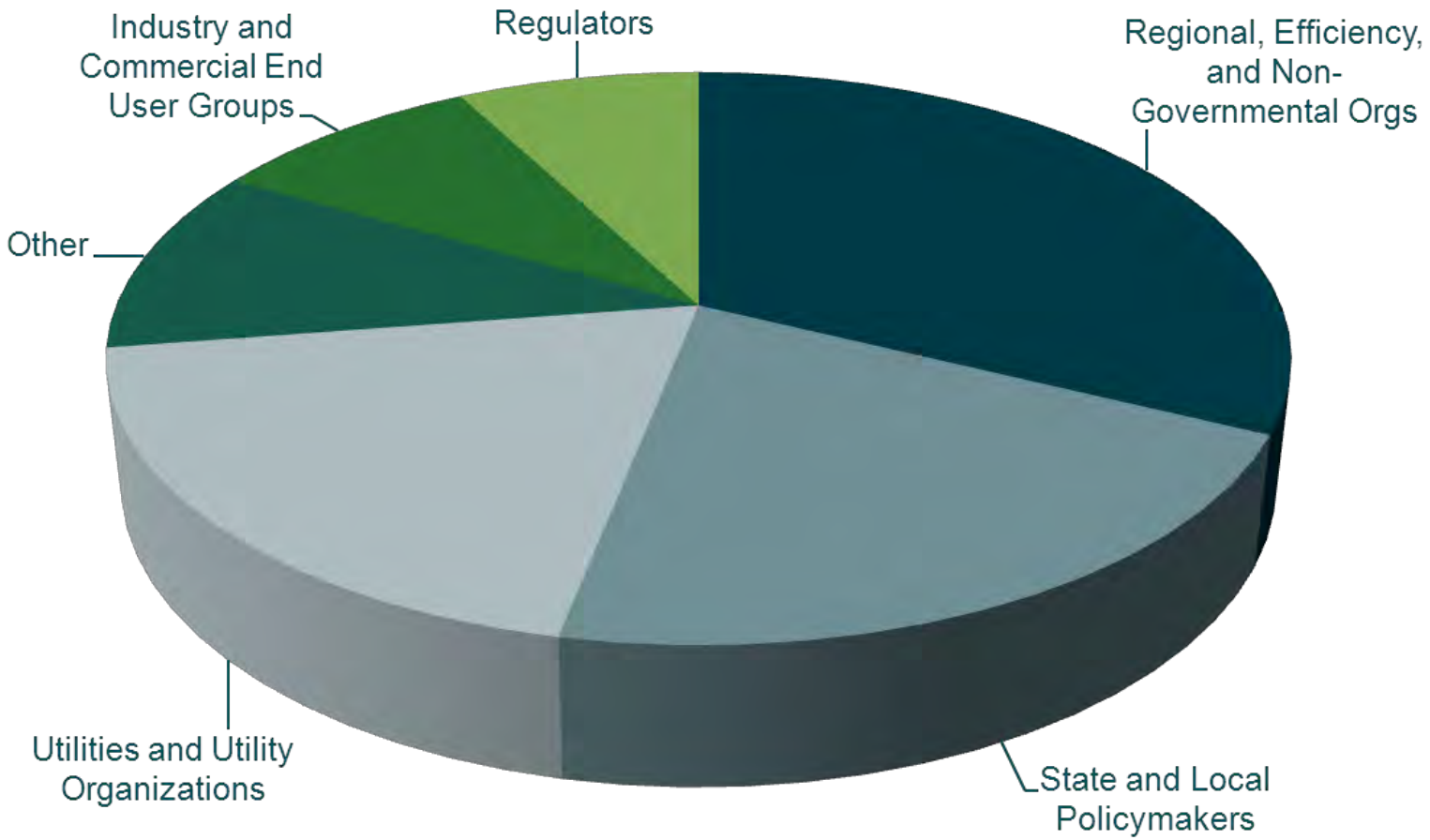
SEE Action Working Groups

Working Group Leadership

Susan Ackerman, OR PUC
Vaughn Clark, OK SEO
Todd Currier, WA SEO
Jennifer Easler, IA Consumer Advocate
Joshua Epel, CO PUC
Jim Gallagher, NY ISO
Bryan Garcia, CT Clean Energy Fund
Frank Murray, NY SEO
Pat Oshie, WA PUC
Phyllis Reha, MN PUC
Cheryl Roberto, OH PUC
Janet Streff, MN SEO
Keith Welks, PA Treasury
Malcolm Woolf, MD SEO

Who is SEE Action?

Over 200 professionals representing over 130 organizations



Energy Providers in the SEE Action